

MATERIAL EXCLUSIVO
GRUPO CURSO FÁCIL
CLIQUE AQUI
E ACESSE + MATERIAIS



Informática para Caixa Econômica Federal - CEF (Técnico Bancário)

Prof. Diego Carvalho, Prof. Marcos Vinícius Alves Franco, Prof. Diego Carvalho

13 Análise de Informações - Mineração de Dados

Documento última vez atualizado em 02/01/2024 às 20:15.





Índice

13.1)	Análise de Informações Mineração de Dados Teoria	3
13.2)	Análise de Informações Mineração de Dados - Resumo	91
13.3)	Mapa Mental Análise de Informações Mineração de Dados	100
13.4)	Questões Comentadas Análise de Informações Mineração de Dados Multibancas	101

13.1 Análise de Informações Mineração de Dados Teoria

▶ Para assistir ao vídeo correspondente, acesse o LDI.

Apresentação do Tópico

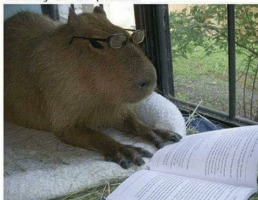
Fala, galera! O assunto do nosso tópico é **Mineração de Dados (Data Mining)**! Pessoal, o que um mineiro faz? Faz pão de queijo e feijão tropeiro, Diego! Não estou me referindo a esse mineiro, zé mané! Estou me referindo àquele que trabalha com mineração! Ele escava grandes áreas de terra em busca de minérios, combustíveis ou até pedras preciosas. O mineiro do mundo moderno escava bases de dados em busca de dados relevantes para uma organização e esse é o tema de hoje :)



PROFESSOR DIEGO CARVALHO - www.instagram.com/professordiegocarvalho [1]



estudando mineração pra
desvendar os mistérios desse teu
coração de pedra



Vainnn



Galera, todos os tópicos da aula possuem Faixas de Incidência, que indicam se o assunto cai muito ou pouco em prova. Diego, se cai pouco para que colocar em aula? Cair pouco não significa que não cairá justamente na sua prova! A ideia aqui é: se você está com pouco tempo e precisa ver somente aquilo que cai mais, você pode filtrar pelas incidências média, alta e altíssima; se você tem tempo sobrando e quer ver tudo, vejam também as incidências baixas e baixíssimas. Fechado?

Mineração de Dados

Contextualização

INCIDÊNCIA EM PROVA: Altíssima

Galera, vamos falar agora sobre dados! *Professor, eu não aguento mais falar sobre dados!* Eu sei, mas fazer o que se esse assunto está na moda: Data Warehouse, Data Mart, Big Data, Data Analytics, Data Privacy, Data Lake, Data Security... e a estrela da nossa aula de hoje: Data Mining! **Também chamada de Mineração de Dados ou Prospecção de Dados, trata-se do processo de explorar grandes quantidades de dados^[2] à procura de padrões consistentes.**

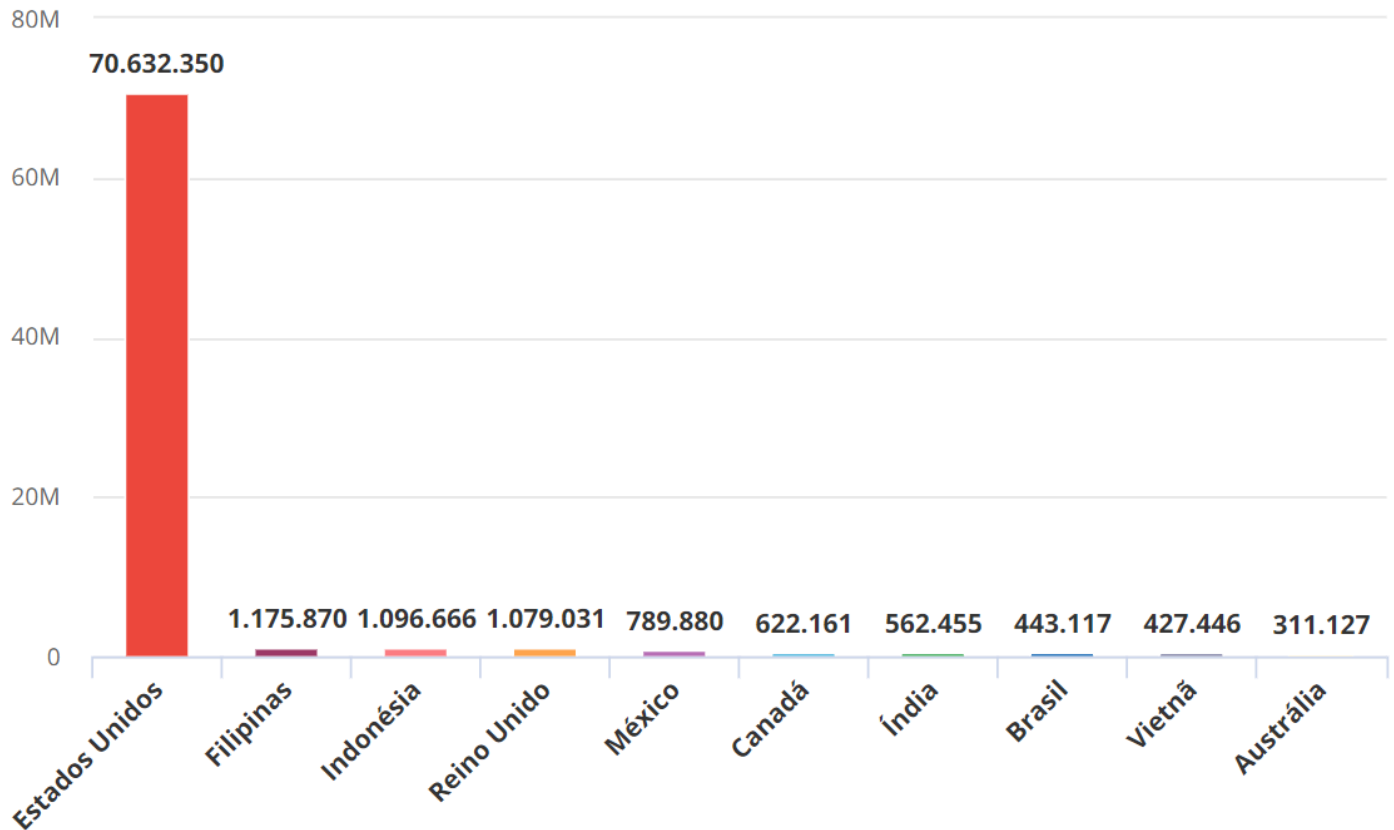
Nada melhor para explicar um conceito do que por meio de exemplos. E o exemplo que eu vou utilizar é do escândalo da Cambridge Analytica. *Que fofoca é essa que eu não fiquei sabendo, Diego?* **Eu vou te contar: uma empresa britânica supostamente especializada em consultoria política conseguiu obter de forma ilegal dados sobre mais de oitenta milhões de usuários do Facebook sem consentimento.**

Esses dados foram utilizados em vários países para verificar o perfil de eleitores e influenciar suas opiniões no intuito de ajudar políticos a vencerem eleições. Por meio de consultas às páginas curtidas, data de nascimento, cidade, sexo, etc, uma aplicação de mineração de dados conseguiu traçar perfis psicológicos dessas pessoas e criar campanhas ou propagandas direcionadas de forma mais eficaz para influenciar em suas convicções políticas.



Essa imagem acima é do Mark Zuckerberg – proprietário do Facebook – prestando esclarecimentos em uma sessão de questionamentos no congresso americano. Ele foi ao Congresso para esclarecer como o Facebook reagiu ao vazamento de dados de 87 milhões de

peças pela consultoria política Cambridge Analytica e como sua empresa de tecnologia trabalha para proteger os dados de seus usuários. *Professor, o Brasil estava na lista?*



Olha lá... talvez seus dados tenham sido analisados pela Cambridge Analytica. *Professor, essa influência resultou em algo?* Bem, há suspeitas de que a empresa britânica teria sido utilizada para ajudar Donald Trump a vencer as eleições presidenciais norte-americanas de 2016, mas isso é outro papo! O que importa aqui é que técnicas similares são utilizadas por empresas como Twitter, Google ou Amazon para descobrir o que nós queremos ver ou comprar.

E isso não é utilizado apenas em anúncios e política: a mineração de dados possibilita que companhias aéreas prevejam quem perderá um voo; é capaz de informar a grandes lojas de departamento quem possivelmente está grávida; ajuda médicos a identificarem infecções fatais; e impressionantemente podem ser utilizadas até para prever – por meio de dados celulares – possíveis atentados terroristas em diversos países.

O poder da mineração de dados e todo o enfoque que está tendo atualmente podem fazer parecer que é uma varinha mágica capaz de salvar uma empresa ou destruir uma democracia. Não é nada disso: é simplesmente a aplicação de técnicas estatísticas capazes de fazer uma varredura em uma quantidade massiva de dados em busca de padrões impossíveis de serem detectados por seres humanos.

Esses padrões não são baseados na intuição humana, mas no que os dados nos sugerem. Isso pode parecer uma ideia revolucionária, mas nada mais é do que a previsão do tempo faz há décadas (aliás, a mineração de dados é muito parecida com a meteorologia). Os meteorologistas basicamente buscam duas coisas: primeiro, eles querem descrever padrões

climáticos genéricos; segundo, eles querem prever a temperatura e umidade de um dia qualquer.

De forma similar, os cientistas de dados do Spotify podem estar interessados em descobrir o perfil de pessoas que curtem rock pesado e sugerir músicas de gêneros similares. O grande lance aqui é descrever e prever algo não através de um estudo sociológico realizado por grandes especialistas em pesquisas de mestrado, por exemplo, mas simplesmente analisar uma quantidade massiva de dados. *Bacana?*

E como isso poderia ser feito no Spotify? Poderiam ser analisados padrões de gravadoras, reviews publicados na Internet, parcerias entre músicos, idade, localização, entre outros. A mineração de dados é mais sobre identificar padrões do que sobre explicá-los. O que isso significa? Isso significa que muitas vezes você encontrará um padrão específico, mas não fará ideia do que ele significa. Vejam que legal...

Vou contar para vocês uma lenda que existe desde a década de noventa e que eu ouvi quando estava na faculdade: fraldas x cervejas! Você vai ao mercado comprar uma pasta de dente e do lado tem um... fio dental! Você aproveita e decide comprar também um espaguete e do lado tem um... molho de tomate! *Professor, não há nada de interessante nisso – você coloca produtos que geralmente são utilizados juntos um ao lado do outro.*

É verdade, mas e se eu chego em um supermercado e, ao lado da seção de fraldas, existe uma pilha de caixas de cerveja? Olhando assim não há nenhuma lógica, mas vamos imaginar o seguinte cenário: João é casado com Maria! Ambos passaram em um concurso público, adquiriram uma estabilidade financeira e tiveram um filho, que é recém-nascido). Um dia, João vem voando do trabalho para fazer um esquentado antes do jogo do maior time da história: **Flamengo.**

Ele chega em casa, toma um banho e vai correndo para frente da televisão para assistir o Mengão dar aquela surra tradicional no adversário. Enquanto isso, Maria vai ao supermercado, depois busca o filho na creche e, por fim, volta para casa cheia de sacolas e o bebê! Só que ao retirar os itens das sacolas, ela percebe que esqueceu de comprar as fraldas. Ela diz para João: *“Amor, vá ao supermercado comprar fraldas porque eu esqueci de comprar e não temos mais, por favor”.*

João argumenta dizendo que faltam vinte minutos para começar o jogo, mas não convence sua esposa. O que ele faz? Pega o carro e sai voado até o supermercado mais próximo. Só que ele chega na seção onde se encontram as fraldas e percebe que do lado tem uma pilha de caixas de cervejas. Ele pensa: *“Poxa! Hoje é quarta-feira, trabalhei o dia todo, estou cansado, meu time vai jogar, eu estou aqui comprando fralda... eu acho que mereço tomar uma cervejinha!”*

E João volta feliz e contente para assistir ao jogo! *Pessoal, por que essa história é, na verdade, uma lenda? Porque o Flamengo não ganha de ninguém, professor!* Que isso? Não façam piadas com meu Mengão! Galera, essa história é uma lenda porque ela não é totalmente verdade! **Uma farmácia¹ [3] americana realmente identificou essa relação, mas não foi baseado na**

análise de dados genérica, foi apenas uma pessoa que ficou curiosa e decidiu pesquisar exatamente essa relação.



Por outro lado, a ideia aqui é mostrar que essa correlação seria possível por meio de ferramentas de mineração de dados. E, retornando a ideia anterior, em muitos casos você encontrará correlações que são acidentais e em outros casos você encontrará correlações, mas não conseguirá explicá-las – como vimos no caso das fraldas e cervejas. **Uma rede varejista, por exemplo, descobriu que a venda de colírios aumentava na véspera de feriados.** Por que, professor?

Ninguém sabe! Ainda não encontraram uma resposta para isso, de todo modo essa rede passou a preparar seus estoques e promoções com base nesse cenário. A Sprint, uma das líderes no mercado americano de telefonia, desenvolveu – com base no seu armazém de dados – um método capaz de prever com 61% de confiança se um consumidor trocaria de companhia telefônica dentro de um período de dois meses.

Com um marketing agressivo, ela conseguiu evitar o cancelamento de 120.000 clientes e uma perda de 35 milhões de dólares em faturamento. *Professor, dá um exemplo brasileiro aí?* Claro, tem um exemplo interessantíssimo da SEFAZ-AM! **Com o uso de mineração de dados, foram verificadas correlações entre o estado civil e salários se servidores da Secretaria de Fazenda do Estado do Amazonas.** Como assim, professor?

Notou-se que cerca de 80% dos servidores de maior poder aquisitivo deste órgão eram divorciados, enquanto em outras instituições (Ex: Secretaria de Educação) esta média de divorciados era inferior a 30%. *Qual seria a explicação para esse fenômeno?* **Os dados parecem sugerir que servidores com maior poder aquisitivo se envolvem mais em relações extraconjugais, resultando geralmente no término do casamento.**

Como na Secretaria de Fazenda do Amazonas há servidores geralmente com maior poder aquisitivo do que na Secretaria de Educação (que é composta basicamente por professores), a explicação parece válida. **Outra interpretação possível seria a de que, com poder aquisitivo individual mais elevado, não haveria razão para manter um casamento que não estivesse feliz.** Logo, prevalece a máxima de que é melhor só do que mal acompanhado...

Definições e Terminologias

INCIDÊNCIA EM PROVA: Altíssima

O termo Mineração de Dados remonta à década de 1980, quando seu objetivo era extrair conhecimento dos dados. **Nesse contexto, o conhecimento é definido como padrões interessantes que são geralmente válidos, novos, úteis e compreensíveis para os seres humanos.** Se os padrões extraídos são interessantes ou não, depende de cada aplicação específica e precisa ser verificado pelos especialistas dessas aplicações.

Já o termo Análise de Dados/Informações tornou-se popular no início dos anos 2000. **A análise de dados é definida como a aplicação de softwares para a análise de grandes conjuntos de dados para o suporte de decisões.** A análise de dados é um campo muito interdisciplinar que adotou aspectos de muitas outras disciplinas científicas, como estatística, teoria de sinais, reconhecimento de padrões, inteligência computacional, aprendizado de máquina e pesquisa operacional.

A Análise de Dados é uma ferramenta importante para entender tendências e projeções, permitindo identificar padrões, correlações e relacionamentos entre diferentes pontos de dados e fornecer informações sobre os dados que podem ser usados para fazer previsões e projeções. **A análise de dados pode ser usada para analisar tendências passadas e fazer projeções sobre tendências futuras, bem como para identificar áreas de melhoria potencial e áreas de risco.**

Ela também pode ser usada para avaliar a eficácia de uma estratégia ou para identificar áreas de melhoria. Ao analisar tendências e projeções, as empresas podem entender melhor seus mercados e clientes e tomar melhores decisões sobre como posicionar seus produtos e serviços. Mais recentemente ainda, surgiram com mais uma palavra da moda: ***Analytics!*** Trata-se do processo sistemático de coletar, analisar e interpretar dados a fim de obter insights e tomar decisões.

Honestamente, isso é tudo jogada de marketing. De tempos em tempos, alguém cria uma palavra da moda para revolucionar o mundo de tecnologia da informação, mas é tudo quase a mesma coisa. **Logo, nós vamos nos ater ao que efetivamente mais cai em prova: Data Mining (Mineração de Dados).** É importantíssimo saber esse conceito, visto que as bancas adoram cobrar maneiras diferentes de se definir mineração de dados. Vamos lá...

DEFINIÇÕES DE DATA MINING

Data Mining é o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida.

Palavras-chave: exploração; informação implícita desconhecida.

Data Mining é uso de teorias, métodos, processos e tecnologias para organizar uma grande quantidade de dados brutos para identificar padrões de comportamentos em determinados públicos.

Palavras-chave: teorias; métodos; processos; tecnologias; organizar dados brutos; padrões de comportamentos.

Data Mining é a categoria de ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados.

Palavras-chave: ferramenta de análise; open-end; tendências e padrões.

Data Mining é o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

Palavras-chave: descoberta; correlações; padrões; tendências; reconhecimento de padrões; estatística; matemática.

Data Mining constitui em uma técnica para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação.

Palavras-chave: exploração e análise de dados; padrões; regras; ocultos.

Data Mining é o conjunto de ferramentas que permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas

avançadas de computação como redes neurais, algoritmos genéticos e lógica nebulosa (fuzzy), dentre outras.

Palavras-chave: tendências; padrões; redes neurais; algoritmos genéticos; lógica nebulosa.

Data Mining é o conjunto de ferramentas e técnicas de mineração de dados que têm por objetivo buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.

Palavras-chave: classificação; agrupamento; clusterização; padrões.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes com o intuito de detectar relacionamentos sistemáticos entre variáveis e novos subconjuntos de dados.

Palavras-chave: padrões; relacionamentos.

Data Mining consiste em explorar um conjunto de dados visando a extrair ou a ajudar a evidenciar padrões, como regras de associação ou sequências temporais, para detectar relacionamentos entre estes.

Palavras-chave: exploração; padrões; regras; associação; sequência temporal; detecção.

Data Mining são ferramentas que utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (*cluster analysis*), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

Palavras-chave: estatística; análise de conglomerados; agrupamento.

Data Mining é o conjunto de técnicas que, envolvendo métodos matemáticos e estatísticos, algoritmos e princípios de inteligência artificial, tem o objetivo de descobrir relacionamentos significativos entre dados armazenados em repositórios de grandes volumes e concluir sobre padrões de comportamento de clientes de uma organização.

Palavras-chave: métodos matemáticos e estatístico; inteligência artificial; relacionamentos; padrões; comportamentos.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

Palavras-chave: padrões; regras de associação; sequências temporais; relacionamentos.

Data Mining é o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

Palavras-chave: padrões; utilidade.

Data Mining é um método computacional que permite extrair informações a partir de grande quantidade de dados.

Palavras-chave: extração.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais.

Palavras-chave: exploração; padrões consistentes; regras de associação; sequência temporal.

Data Mining é o processo de analisar de maneira semi-automática grandes bancos de dados para encontrar padrões úteis.

Palavras-chave: padrões.

*Professor, não entendi a utilidade de ver tantas definições diferentes. Galera, eu coloquei todas as definições porque TODAS elas caíram em prova (sem exceção!). Eu retirei todas essas definições de provas anteriores. **Notem como é importante saber de forma abrangente as possíveis definições de um conceito teórico importante.** Dito isso, vamos tentar condensar todos esses conceitos em uma grande definição a seguir:*

Data Mining – Mineração de Dados – é um conjunto de processos, métodos, teorias, ferramentas e tecnologias open-end utilizadas para explorar, organizar e analisar de

forma automática ou semi-automática uma [4] grande quantidade de dados brutos com o intuito de identificar, descobrir, extrair, classificar e agrupar informações implícitas desconhecidas, além de avaliar correlações, tendências e padrões consistentes de comportamento potencialmente úteis – como regras de associação ou sequências temporais – de forma não-trivial por meio de técnicas estatísticas e matemáticas, como redes neurais, algoritmos genéticos, inteligência artificial, lógica nebulosa, análise de conglomerados (clusters), entre outros.

Pronto! Essa é uma definição extremamente abrangente com diversas palavras-chave que remetem à mineração de dados. Sabendo isso, vocês já saberão responder a grande maioria das questões de prova. É importante mencionar também que – apesar de geralmente ser utilizada em conjunto com Data Warehouses ³ [5] – não é obrigatório que o seja! Você pode aplicar técnicas de mineração em diversos outros contextos (inclusive bases de dados transacionais).

É importante mencionar que a mineração de dados necessita, por vezes, utilizar processamento paralelo para dar conta da imensa quantidade de dados a serem analisados. Aliás, as ferramentas de mineração geralmente utilizam uma arquitetura cliente/servidor ou até uma arquitetura web. Outra pergunta que sempre me fazem é se é necessário ser um grande analista de dados para utilizar essas ferramentas de mineração. A resposta é: Não!

Usuários podem fazer pesquisas sem necessariamente saber detalhes da tecnologia, programação ou estatística. Por fim, a mineração de dados pode ser aplicada a uma grande variedade de contextos de tomada de decisão de negócios a fim de obter vantagens competitivas estratégicas. **Em particular, algumas áreas de ganhos significativos devem incluir as seguintes: marketing, finanças, manufatura e saúde.**

CARACTERÍSTICAS de MINERAÇÃO DE DADOS

Há diferentes tipos de mineração de dados: (1) diagnóstica, utilizada para entender os dados e/ou encontrar causas de problemas; (2) preditiva, utilizada para antecipar comportamentos futuros.

As provas vão insistir em afirmar que a mineração de dados só pode ocorrer em bancos de dados muito grandes como Data Warehouses, mas isso é falso – apesar de comum, não é obrigatório.

Em geral, ferramentas de mineração de dados utilizam uma arquitetura web cliente/servidor, sendo possível realizar inclusive a mineração de dados de bases de dados não estruturadas.

Não é necessário ter conhecimentos de programação para realizar consultas, visto que existem ferramentas especializadas que auxiliam o usuário final de negócio.

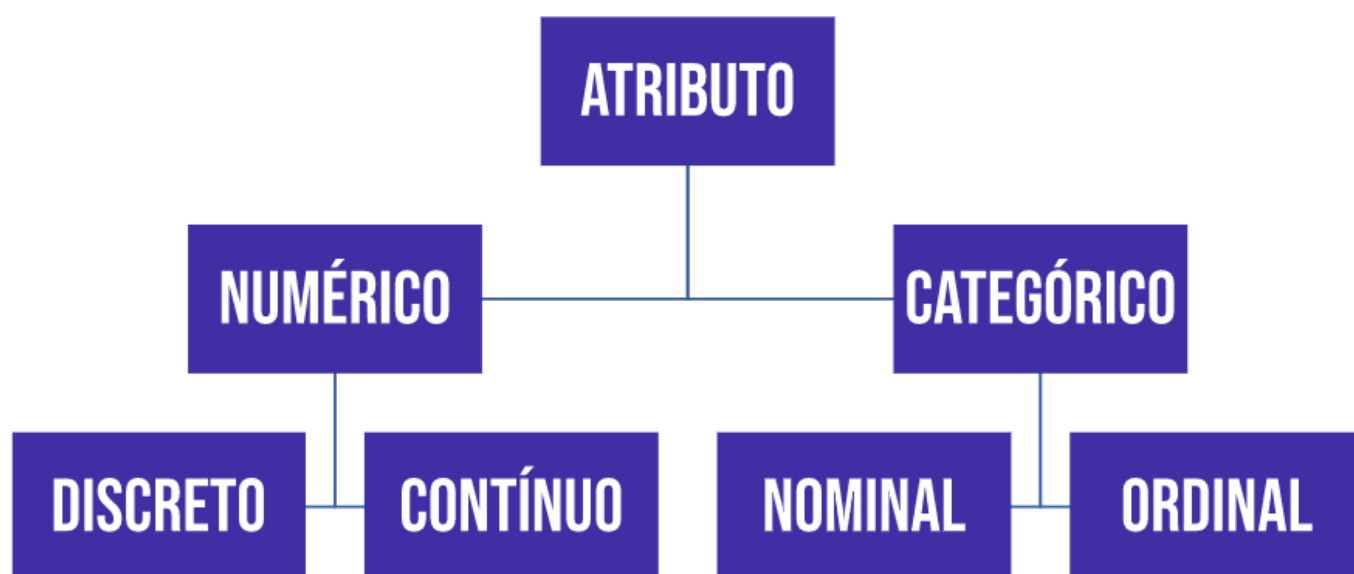
Antes de seguir, vamos ver algumas definições que serão importantes para o entendimento da aula. Primeiro, vamos tratar da classificação de dados quanto à estrutura:

TIPOS DE DADOS	DESCRIÇÃO
DADOS ESTRUTURADOS	São aqueles que residem em campos fixos de um arquivo (Ex: tabela, planilha ou banco de dados) e que dependem da criação de um modelo de dados, isto é, uma descrição dos objetos juntamente com as suas propriedades e relações.
DADOS SEMIESTRUTURADOS	São aqueles que não possuem uma estrutura completa de um modelo de dados, mas também não é totalmente desestruturado. Em geral, são utilizados marcadores (<i>tags</i>) para identificar certos elementos dos dados, mas a estrutura não é rígida.
DADOS NÃO ESTRUTURADOS	São aqueles que não possuem um modelo de dados, que não está organizado de uma maneira predefinida ou que não reside em locais definidos. Eles costumam ser de difícil indexação, acesso e análise (Ex: imagens, vídeos, sons, textos livres, etc).

Nessa aula, vamos nos focar mais nos dados estruturados, isto é, aqueles em que cada linha de uma tabela geralmente corresponde a um objeto (também chamado de exemplo, instância, registro ou vetor de entrada) e cada coluna corresponde a um atributo (também chamado de característica, variável ou *feature*). Nesse contexto, podemos ver na tabela seguinte uma classificação de atributos quanto à sua dependência:

ATRIBUTO DEPENDENTE	Representa um atributo de saída que desejamos manipular em um experimento de dados (também chamado de variável alvo ou variável target).
ATRIBUTO INDEPENDENTE	Representa um atributo de entrada que desejamos registrar ou medir em um experimento de dados.

É muito parecido com uma função matemática, por exemplo: $y = ax + b$. Nesse caso, x seria a variável de entrada (independente) e y seria a variável de saída (dependente).



Por fim, podemos ver no diagrama acima e na tabela abaixo como os atributos se classificam em relação ao seu valor:

TIPOS DE ATRIBUTOS	DESCRIÇÃO
ATRIBUTO NUMÉRICO	Também chamado de atributo quantitativo, é aquele que pode ser medido em uma escala quantitativa, ou seja, apresenta valores numéricos que fazem sentido.
discreto	Os valores representam um conjunto finito ou enumerável de números, e que resultam de uma contagem (Ex: número de filhos, número de bactérias por amostra, número de logins em uma página web, entre outros).
contínuo	Os valores pertencem a um intervalo de números reais e representam uma mensuração (Ex: altura de uma pessoa, peso de uma marmita, salário de um servidor público, entre outros).

TIPOS DE ATRIBUTOS	DESCRIÇÃO
ATRIBUTO categórico	Também chamado de atributo qualitativo, é aquele que pode assumir valores categóricos, isto é, representam uma classificação.
Nominal	São aquelas em que não existe uma ordenação própria entre as categorias (Ex: sexo, cor dos olhos, fumante/não fumante, país de origem, profissão, religião, raça, time de futebol, entre outros).
Ordinal	São aquelas em que existe uma ordenação própria entre as categorias (Ex: Escolaridade (1º, 2º, 3º Graus), Estágio de Doença (Inicial, Intermediário, Terminal), Classe Social (Classe Baixa, Classe Média, Classe Alta), entre outros)

TIPOS DE DADOS	DESCRIÇÃO
DADOS ESTRUTURADOS	São aqueles que residem em campos fixos de um arquivo (Ex: tabela, planilha ou banco de dados) e que dependem da criação de um modelo de dados, isto é, uma descrição dos objetos juntamente com as suas propriedades e relações.
DADOS SEMIESTRUTURADOS	São aqueles que não possuem uma estrutura completa de um modelo de dados, mas também não é totalmente desestruturado. Em geral, são utilizados marcadores (<i>tags</i>) para identificar certos elementos dos dados, mas a estrutura não é rígida.
DADOS NÃO ESTRUTURADOS	São aqueles que não possuem um modelo de dados, que não está organizado de uma maneira predefinida ou que não reside em locais definidos. Eles costumam ser de difícil indexação, acesso e análise (Ex: imagens, vídeos, sons, textos livres, etc).

Nessa aula, vamos nos focar mais nos dados estruturados, isto é, aqueles em que cada linha de uma tabela geralmente corresponde a um objeto (também chamado de exemplo, instância, registro ou vetor de entrada) e cada coluna corresponde a um atributo (também chamado de característica, variável ou *feature*). Nesse contexto, podemos ver na tabela seguinte uma classificação de atributos quanto à sua dependência:

ATRIBUTO DEPENDENTE	Representa um atributo de saída que desejamos manipular em um experimento de dados (também chamado de variável alvo ou variável target).
ATRIBUTO INDEPENDENTE	Representa um atributo de entrada que desejamos registrar ou medir em um experimento de dados.

Processo de Descoberta de Conhecimento

INCIDÊNCIA EM PROVA: ALTA

A Mineração de Dados faz parte de um processo muito maior de descoberta de conhecimento chamada KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Bancos de Dados). O processo de descoberta de conhecimento compreende cinco fases: (1) Seleção; (2) Pré-processamento; (3) Transformação; (4) **Data Mining**; (5) Interpretação e Avaliação – alguns autores possuem uma classificação um pouco diferente como veremos adiante.

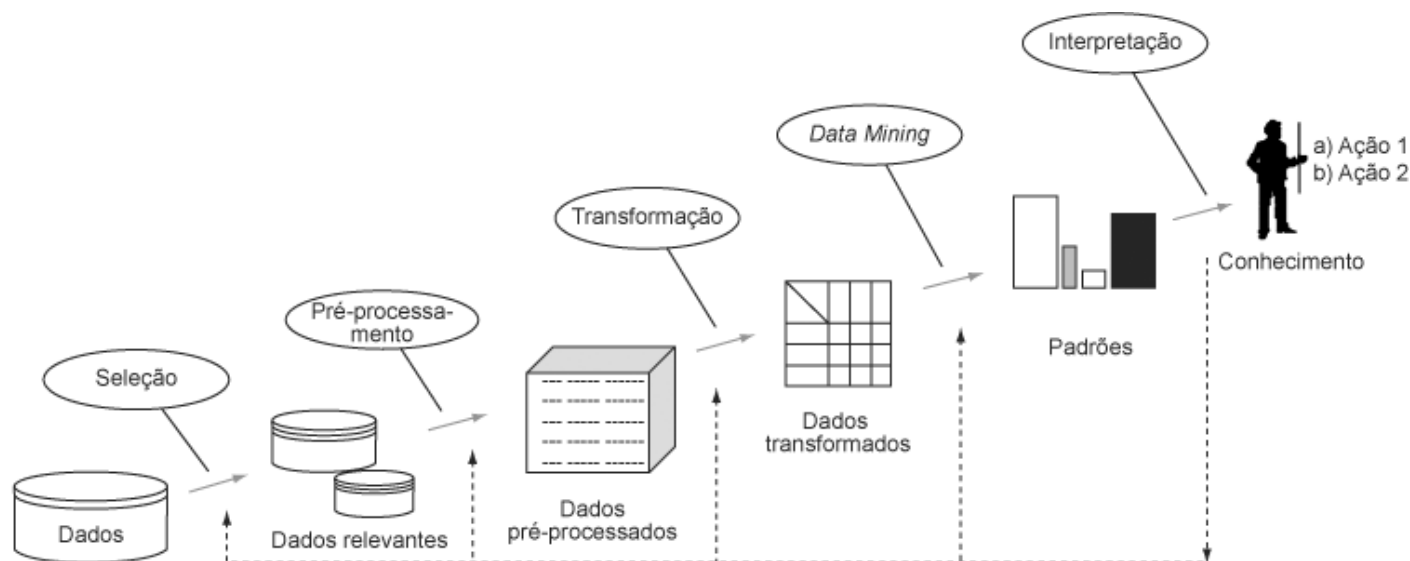


Figura 1. Etapas do processo KDD (Fayyad et al. (1996)).

O processo de descoberta de conhecimento é interativo e iterativo, envolvendo várias etapas com muitas decisões tomadas pelo usuário. É necessário desenvolver uma compreensão do domínio de aplicação e os conhecimentos anteriores relevantes. Dessa forma, a primeira etapa é selecionar um conjunto de dados de diversas bases – ou se concentrar em um subconjunto de variáveis ou amostras de dados – no qual a descoberta será realizada.

Com a seleção de dados relevantes, a segunda etapa será o pré-processamento dos dados. Operações básicas incluem limpeza, remoção de erros, eliminação de redundância, decidir

estratégias para lidar com campos de dados ausentes, entre outros. Com os dados pré-processados, passamos à etapa de transformação, em que os dados são enriquecidos e consolidados em formas apropriadas à mineração, resumizando-os ou agregando-os.

Com os dados transformados, passamos à etapa de mineração de dados. Utilizam-se algoritmos e técnicas para extrair possíveis padrões úteis de dados. Por fim, teremos a descoberta de diversos padrões que serão interpretados e avaliados em busca de padrões realmente interessantes e úteis, além de suas possíveis explicações ou interpretações. Vamos detalhar agora um pouco mais a fase de pré-processamento...

Quando temos bases de dados muito grandes e heterogêneas, é comum termos registros que comprometem a qualidade dos dados, como – por exemplo – registros inconsistentes, falta de informação, registros duplicados, valores discrepantes, entre outros. **No entanto, existem diversas técnicas para pré-processar dados com o intuito de preparar os dados brutos para serem analisados sem erros de incompletudes, inconsistências, ruídos, entre outros**⁴ [6]

A técnica de pré-processamento possui alguns objetivos principais: **melhorar a qualidade dos dados; diminuir a ambiguidade das expressões linguísticas; diminuir a quantidade de dados a ser processado; estruturar as informações como tuplas; e melhorar a eficiência da mineração de dados.** Eslmari Navathe – renomado autor de bancos de dados – considera que o KDD contempla seis etapas:

1	SELEÇÃO DE DADOS	Dados sobre itens ou categorias são selecionados.
2	LIMPEZA DE DADOS	Dados são corrigidos ou eliminados dados incorretos.
3	ENRIQUECIMENTO DE DADOS	Dados são melhorados com fontes de informações adicionais.
4	TRANSFORMAÇÃO DE DADOS	Dados são reduzidos por meio de sumarizações, agregações e discretizações ⁵ [7].
5	MINERAÇÃO DE DADOS	Padrões úteis são descobertos.
6	EXIBIÇÃO DE DADOS	Informações descobertas são exibidas ou relatórios são construídos.

Dito isso, esse autor considera que as quatro primeiras etapas podem ser agrupadas em uma única etapa de pré-processamento. *Fechou? Fechado...*

Principais Objetivos

INCIDÊNCIA EM PROVA: média

Segundo Navathe, a Mineração de Dados costuma ser executada com alguns objetivos finais ou aplicações. De um modo geral, esses objetivos se encontram nas seguintes classes: Previsão, Identificação, Classificação ou Otimização. *Como assim, Diego?* Cara, isso significa que você pode utilizar a mineração de dados com o objetivo que podem ser divididos nas seguintes classes: previsão, identificação, classificação e otimização. Fiquem com o mnemônico para memorizar:

PiCO

Previsão identificação classificação otimização

Objetivo	Descrição
PREVISÃO	A mineração de dados pode mostrar como certos atributos dos dados se comportarão no futuro. Um de seus objetivos é prever comportamentos futuros baseado em comportamentos passados. Exemplo: análise de transações de compras passadas para prever o que os consumidores comprarão futuramente sob certos descontos, quanto volume de vendas uma loja gerará em determinado período e se a exclusão de uma linha de produtos gerará mais lucros. Em tais aplicações, a lógica de negócios é usada junto com a mineração de dados. Em um contexto científico, certos padrões de onda sísmica podem prever um terremoto com alta probabilidade.
IDENTIFICAÇÃO	Padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade. Por exemplo: intrusos tentando quebrar um sistema podem ser identificados pelos programas por eles executados, arquivos por eles acessados ou pelo tempo de CPU por sessão aberta. Em aplicações biológicas, a existência de um gene pode ser identificada por sequências específicas de nucleotídeos em uma cadeia de DNA. A área conhecida como autenticação é uma forma de identificação. Ela confirma se um usuário é realmente um usuário específico ou de uma classe autorizada, e envolve uma comparação de parâmetros, imagens ou sinais contra um banco de dados.

CLASSIFICAÇÃO

A mineração de dados pode particionar os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros. Por exemplo: os clientes em um supermercado podem ser categorizados em compradores que buscam desconto, compradores com pressa, compradores regulares leais, compradores ligados a marcas conhecidas e compradores eventuais. Essa classificação pode ser usada em diferentes análises de transações de compra de cliente como uma atividade pós-mineração. Às vezes, a classificação baseada em conhecimento de domínio comum é utilizada como uma entrada para decompor o problema de mineração e torná-lo mais simples (Ex: alimentos saudáveis, alimentos de festa ou alimentos de lanche escolar são categorias distintas nos negócios do supermercado. Faz sentido analisar o relacionamento dentro e entre categorias como problemas separados). Essa categorização pode servir para codificar os dados corretamente antes de submetê-los a mais mineração de dados.

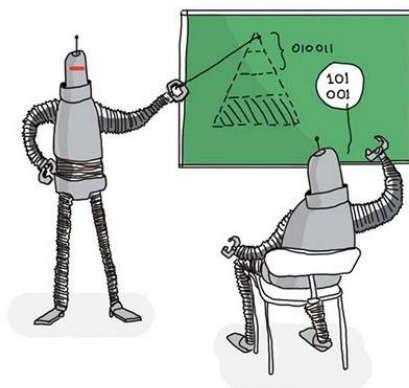
OTIMIZAÇÃO

Um objetivo relevante da mineração de dados pode ser otimizar o uso de recursos limitados, como tempo, espaço, dinheiro ou materiais e maximizar variáveis de saída como vendas ou lucros sob determinado conjunto de restrições. Como tal, esse objetivo da mineração de dados é semelhante à função objetiva, usada em problemas de pesquisa operacional, que lida com otimização sob restrições.

Aprendizado de Máquina

INCIDÊNCIA EM PROVA: baixíssima

Aprendizado de Máquina (do inglês, *Machine Learning*) é a área da inteligência artificial que busca desenvolver técnicas computacionais sobre aprendizado assim como a construção de sistemas capazes de adquirir conhecimento de forma autônoma que tome decisões baseado em experiências acumuladas por meio da solução bem-sucedida de problemas anteriores. **Trata-se de uma ferramenta poderosa para a aquisição automática de conhecimento por meio da imitação do comportamento de aprendizagem humano com foco em aprender a reconhecer padrões complexos e tomar decisões.**



Vamos ver exemplo bem simples: imagine que você deseja criar um software para diferenciar maçãs e laranjas. Você tem dados que indicam que as laranjas pesam entre 150-200g e as maçãs entre 100-130g. Além disso, as laranjas têm cascas ásperas e as maçãs têm cascas suaves (que você pode representar como 0 ou 1). **Ora, se você tem uma fruta que pesa 115g e possui casca suave, seu programa pode determinar que provavelmente se trata de uma maçã.**

De outra forma, se a fruta tem 175g e possui casca áspera, provavelmente se trata de uma laranja. Qualquer coisa fora desses limites também não será nem um nem outro. *E se sua fruta tem casca suave, mas pesa apenas 99g?* Você sabe que provavelmente é uma maçã, mas o seu programa não sabe disso. Logo, quanto mais dados você tiver, mais precisos serão os seus resultados. **A parte legal é que seu algoritmo pode ir aprendendo sozinho o que é uma laranja ou uma maçã.**

Cada vez que o usuário confirma que o programa acertou ou errou, ele é capaz de aprender e melhorar! Vejam no exemplo abaixo uma tecnologia que está ficando cada vez mais comum: reconhecimento fácil! Hoje em dia, a China possui mais de 200 milhões de câmeras de vigilância com essa tecnologia e, quanto mais ela for testada, mais a máquina aprenderá e ficará mais precisa. *Sério, tem coisa mais linda que a área de Tecnologia da informação.*



Mineração de Texto (Text Mining)

INCIDÊNCIA EM PROVA: baixíssima

A Mineração de Texto é um meio para encontrar padrões interessantes/úteis em um contexto de informações textuais não estruturadas, combinado com alguma tecnologia de extração e de recuperação da informação, processo de linguagem natural e de sumarização ou indexação de documentos. A internet está cheia de informações e processá-las pode ser uma tarefa e tanto, mas essa tarefa pode ser facilmente executada por meio de ferramentas de *Text Mining*.

Para ser bem honesto, qualquer informação é inútil se não pudermos interpretá-la da maneira adequada. As tecnologias de aprendizado de máquina são usadas para extrair informações ocultas ou abstratas de um pedaço de texto utilizando um computador. **O resultado é inovador porque descobrimos tendências e opiniões inexploradas que têm implicações inacreditáveis em diversos campos do conhecimento.**

Há alguns anos, nós estamos testemunhando uma enorme reviravolta em todos os setores devido à aceitação das mídias sociais como uma plataforma para compartilhar opiniões e comentários sobre qualquer coisa, incluindo grandes marcas, pessoas, produtos, etc. **Os usuários agora utilizam plataformas como o Facebook e o Twitter para compartilhar seus sentimentos. Hoje mesmo eu estava navegando pela minha linha do tempo e vi o seguinte tweet:**



Antonio Tabet ✓
@antoniotabet

Nunca, jamais e em tempo algum comprem qualquer coisa no @privalia_br. Poupe seu tempo do aborrecimento que é consertar os erros deles.
#FicaADica cc: @Privalia_Br_SAC

12:14 AM · 2 de out de 2019 · Twitter Web App

10 Retweets 265 Curtidas



Privalia @privalia_br · 16 h

Em resposta a @antoniotabet e @Privalia_Br_SAC

Oi Antonio, tudo bem? Quero te ajudar! Me informe o número do seu pedido, por favor?

1



1



Antonio Tabet ✓ @antoniotabet · 13 h

82711228

2



Observem que a empresa não é boba e já foi atuar pela própria rede social!

Pois é, as mídias sociais agora estão preenchidas com milhões de avaliações sobre diversos produtos, imagine então sobre uma marca. Dessa forma, torna-se imperativo que qualquer marca utilize alguma ferramenta de mineração de textos para conhecer os sentimentos de seus clientes a fim de atendê-los bem. Com essas ferramentas, um profissional de marketing pode conhecer todas as tendências e opiniões relacionadas aos seus concorrentes em potencial.

*O que as pessoas sentem e dizem está espalhado por toda a Internet, mas quem se daria ao luxo de avaliar e estudar tudo isso? **Vamos ser sinceros: as ferramentas de mineração de texto neste contexto podem ser um potencial enorme de fonte de renda para esse sujeito de marketing!*** Enfim, esse é só um panorama geral para que vocês entendam a abrangência da aplicação dessa tecnologia no mundo hoje. Voltemos à teoria...

Em suma, a mineração de texto tem como objetivo a busca de informações relevantes e a descoberta de conhecimentos significativos a partir de documentos textuais não estruturados ou semiestruturados. **Este processo envolve um grau de dificuldade significativo considerando que as informações normalmente estão disponíveis em linguagem natural, sem a preocupação com a padronização ou com a estruturação dos dados – sua matéria prima é a palavra!**

Por fim, um bom exemplo de Mineração de Texto é o Processamento de Linguagem Natural (PLN). Trata-se de uma área dentro da Inteligência Artificial que busca fazer com que os computadores entendam e simulem uma linguagem humana. **É utilizado em diversas ferramentas como Google Tradutor, Sistemas de Reconhecimento de Falas e Nuvem de Palavras – esse último é um dos que eu acho mais interessantes.**

Uma vem de palavras é um recurso gráfico (usado principalmente na internet) para descrever os termos mais frequentes de um determinado texto. O tamanho da fonte em que a palavra é apresentada é uma função da frequência da palavra no texto: palavras mais frequentes são desenhadas em fontes de tamanho maior, palavras menos frequentes são desenhadas em fontes de tamanho menor. Eu gosto muito de brincar com essas ferramentas, vejam só...

<https://monkeylearn.com/word-cloud/> [8]

Você pode inserir qualquer texto e ele devolverá uma nuvem de palavras com tamanho proporcional a frequência. Por curiosidade, eu inseri a letra de Faroeste Caboclo do Legião Urbana e abaixo está o resultado. A palavra que mais aparece é “Santo Cristo”, depois “Maria Lúcia”, e assim por diante. Claro que se trata de uma ferramenta de inteligência artificial capaz

similares. Conforme é apresentado na imagem a seguir, podemos dividi-las em duas categorias: Preditivas e Descritivas. Vejamos na imagem seguinte...

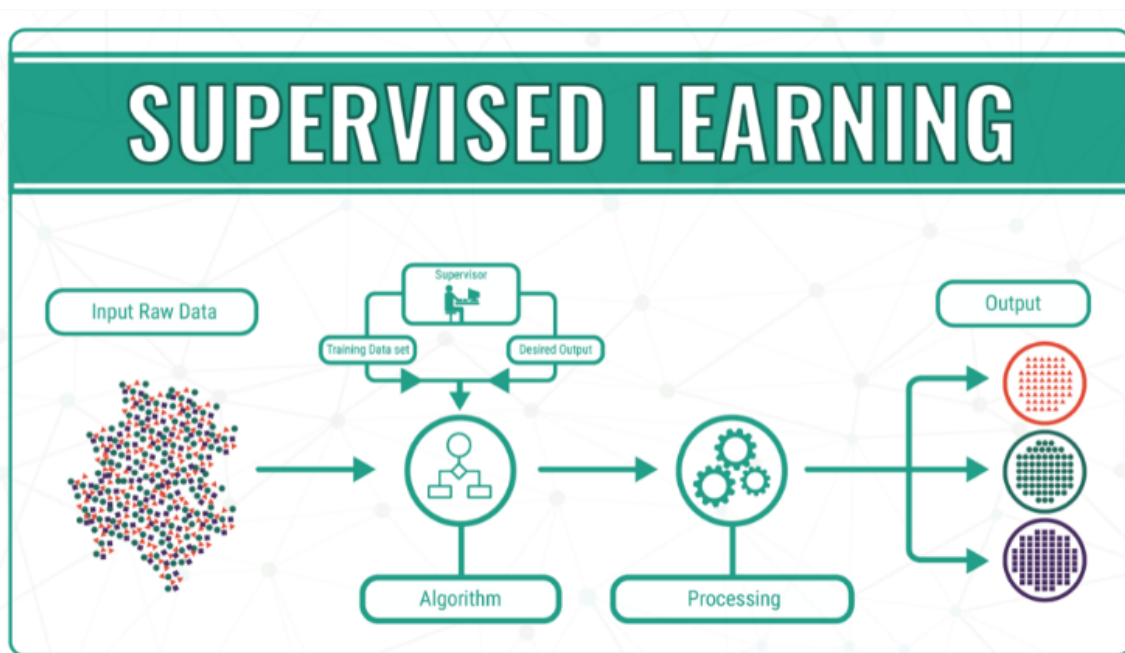


Técnicas Preditivas buscam prever os valores dos dados usando resultados conhecidos coletados de diferentes conjuntos de dados, isto é, busca-se prever o futuro com base dos dados passados. Já **Técnicas Descritivas** buscam descrever relacionamentos entre variáveis e resumir grandes quantidades de dados. Eles usam técnicas estatísticas para encontrar relações entre variáveis, como correlações e associações.

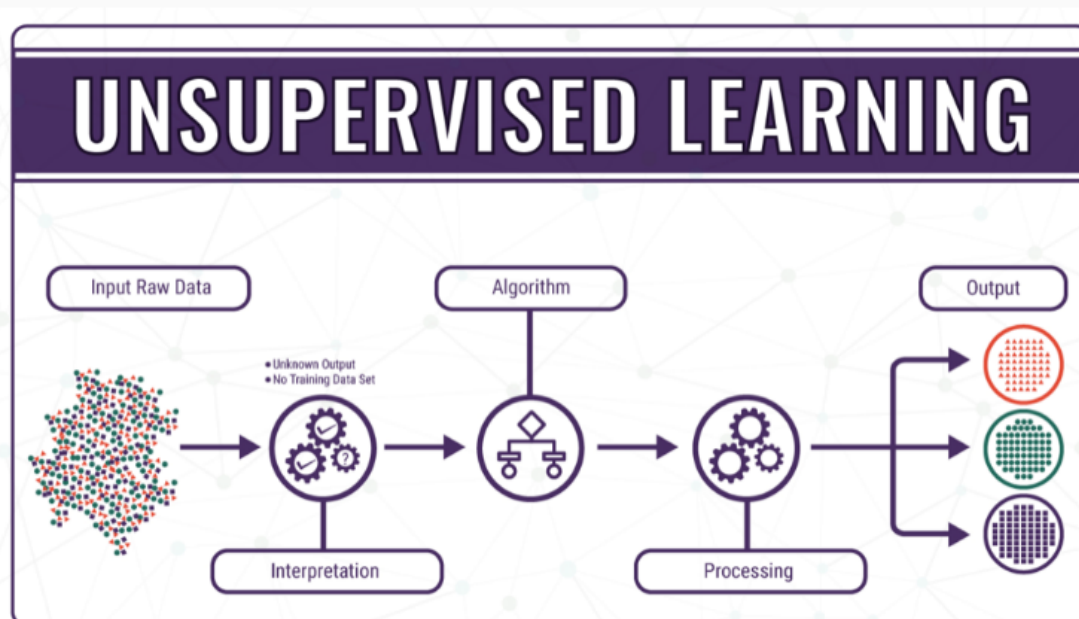
Antes de seguir para o detalhamento de cada uma das técnicas, vamos falar sobre o conceito de aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço⁶_[9]. É bem simples: uma técnica de aprendizado supervisionado é aquela que necessita de supervisão ou interação com um ser humano, enquanto uma técnica de aprendizado não supervisionado não necessita desse tipo de supervisão ou interação. *Calma que ficará mais claro...*

No aprendizado supervisionado, um ser humano alimenta o algoritmo com categorias de dados de saída de tal forma que o algoritmo aprenda como classificar os dados de entrada nas categorias de dados de saída pré-definidas. Vejam na imagem seguinte que há um conjunto de pontinhos e eu desejo classificá-los de acordo com as suas cores em vermelho, verde e roxo (é só um exemplo, as categorias poderiam ser pontinhos cujo nome da cor começa com a letra V ou R).

Note que – de antemão – eu já defini quais serão as categorias de saída que desejo, logo o algoritmo irá receber os dados brutos de entrada, irá processá-los e aprenderá a classificá-los em cada uma das categorias de saída que eu defini inicialmente. **Se um ser humano interferiu no algoritmo pré-definindo as categorias de saída do algoritmo, o algoritmo utilizou um aprendizado supervisionado porque ele aprendeu a categoria/rótulo, mas com o auxílio de um ser humano.**



Já no aprendizado não supervisionado, um ser humano não alimenta o algoritmo com categorias de dados de saída pré-definidas. Vejam na imagem anterior que há um conjunto de pontinhos e o algoritmo – por si só – interpreta esses dados de entrada em busca de similaridades, padrões e características em comum, realiza o processamento e ele mesmo os categoriza da maneira que ele acha mais interessante sem nenhuma interferência humana durante o processo.

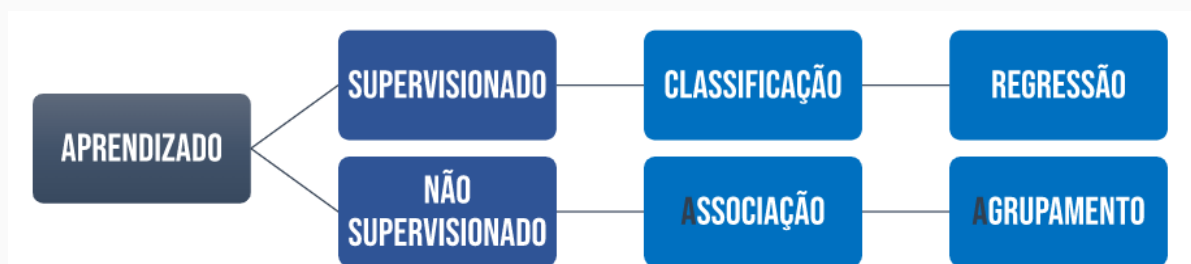


Existe também uma terceira categoria chamada **Aprendizado por Reforço**, que usa recompensas e punições para ajudar os algoritmos a aprender. O princípio é simples: quando um agente de aprendizado realiza uma ação que leva a uma recompensa, ele irá tentar repetir

essa ação para conseguir a recompensa novamente. Por outro lado, quando uma ação leva a uma punição, o agente evitará essa ação no futuro. O objetivo aqui é maximizar a recompensa ao longo do tempo!

Imagine um programa de computador que joga xadrez com um humano. O programa receberia um conjunto de regras e recompensas e, então, tentaria encontrar o melhor movimento a ser feito para maximizar suas recompensas (vencer o adversário). Ao jogar e tomar decisões na tentativa e erro, aprenderia com seus sucessos e insucessos e, eventualmente, se tornaria um jogador de xadrez experiente (tanto que hoje em dia é quase impossível ganhar de um computador).

As técnicas de Classificação e Regressão são consideradas de aprendizado supervisionado; já as técnicas de Associação, Agrupamento são de aprendizado não-supervisionado^[10].

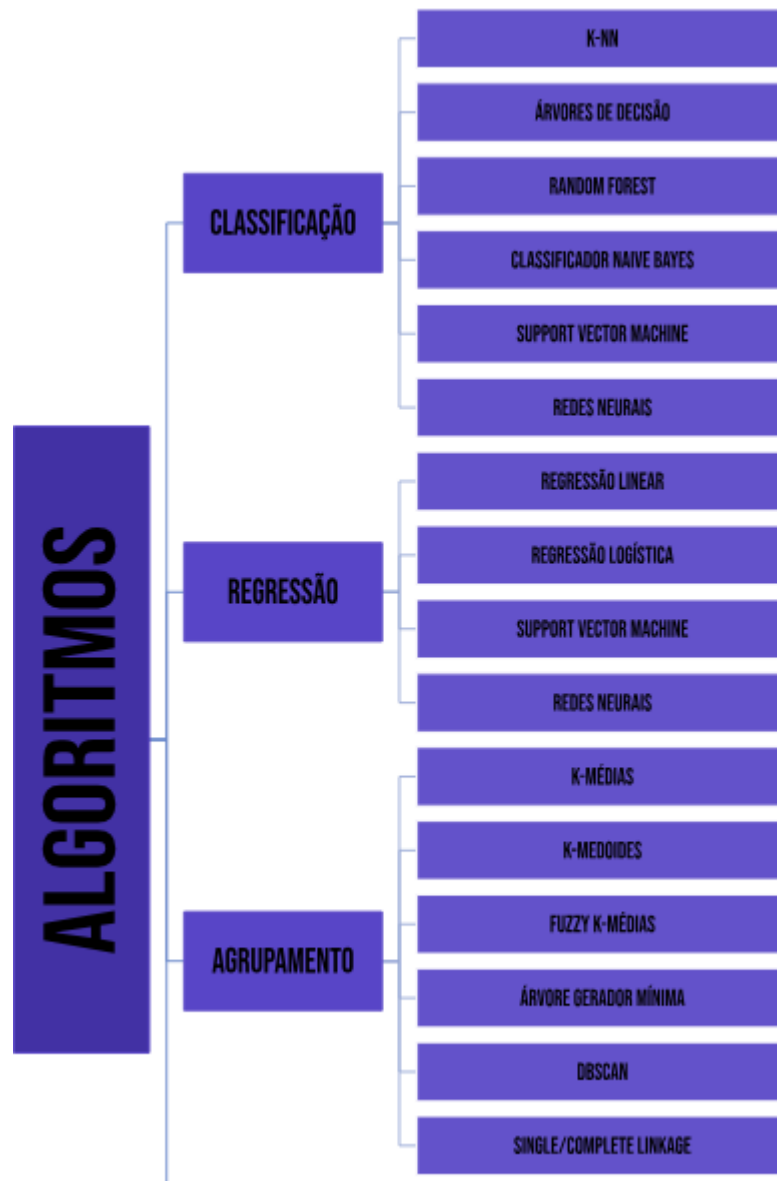


ATENÇÃO MÁXIMA

Galera, agora um ponto importante: nas próximas 50 páginas (mais ou menos), nós veremos essas quatro técnicas/tarefas de mineração de dados e aprendizado de máquina (que possuem um **excelente** custo/benefício) e dentro de cada uma das quatro nós estudaremos diversos algoritmos que implementam de maneiras diferentes essas tarefas (que têm um **péssimo** custo/benefício). Então, eu recomendo que vocês analisem bem o tipo de estudo que desejam...

Existem dezenas de questões sobre as técnicas/tarefas e pouquíssimas questões sobre os algoritmos que as implementam. *Ué, Diego! Então por que você não deixa logo essa parte de fora da aula?* Porque a aula tem que atender tanto aos alunos que desejam fazer um estudo mais aprofundado quanto aos alunos que desejam fazer um estudo mais superficial. Além disso, como esse é um conteúdo relativamente novo, é possível que as provas aprofundem mais com o tempo.

Na próxima página, há um diagrama com os principais algoritmos de implementação das tarefas de mineração de dados e aprendizado de máquina (nem todas serão explicadas nessa aula).



Detecção de Anomalia

INCIDÊNCIA EM PROVA: baixíssima

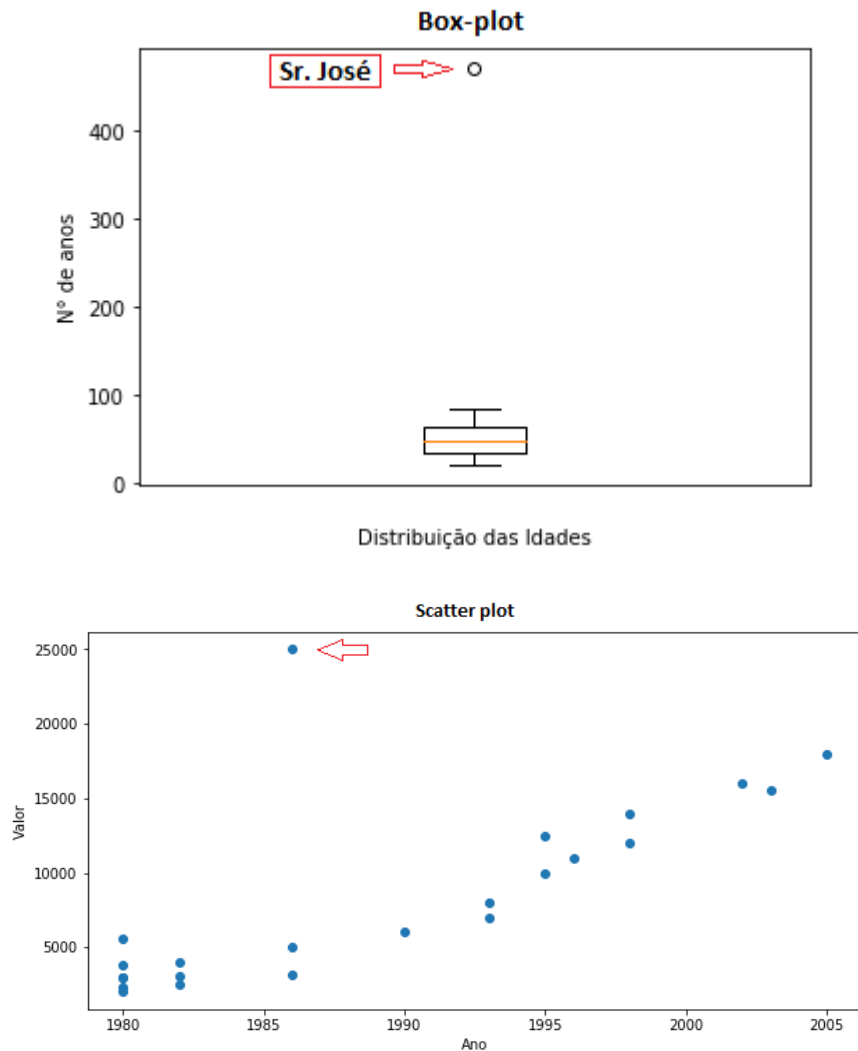
Antes de falarmos especificamente das quatro principais técnicas de mineração, vamos falar sobre a Detecção de Anomalia! Ela também é uma tarefa de mineração, porém pode ser utilizada em conjunto com as outras. **O padrão é que ela funcione com algoritmos de aprendizado não supervisionados, mas ela também pode funcionar com algoritmos de aprendizado supervisionados.**

O que é uma anomalia? Em nosso contexto, é aquele famoso ponto “fora da curva” – também conhecido em inglês como *outlier*. Como assim, professor? Galera, a média de altura para um homem no mundo é de 1,74m! Eu, por exemplo, tenho 1,77m e estou bem próximo da média, no

entanto algumas vezes nós encontramos exceções. Na imagem ao lado, nós temos o nepalês Chandra Dangi e o turco Sultan Kosen – o primeiro é o menor homem do mundo atualmente, medindo 0,54m e o segundo é o maior homem do mundo atualmente, medindo 2,51m. *É comum ver pessoas desses tamanhos?* Não! Por essa razão, esses dados de altura são considerados anomalias, exceções, aberrações justamente porque eles fogem drasticamente da normalidade e do padrão esperado. *E qual é o problema disso?* **O problema é que isso pode causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise.** *E o que causa uma anomalia?*



Bem, a causa pode ter uma origem natural ou artificial. **Uma pessoa que há trinta anos declara ter crescimento de patrimônio anual de 1% e, de repente, declara ter tido um crescimento de 1000% pode ter ganhado na megasena – essa é uma anomalia natural.** Por outro lado, essa mesma pessoa pode simplesmente ter errado na hora de digitar e declarou patrimônio de R\$10.000.000 em vez de R\$100.000 – essa é uma anomalia artificial.



Anomalias artificiais podem partir de erros de amostragem, erros de processamento de dados, erros na entrada de dados, erros de medida ou erros intencionais. **Bem, uma forma de detectar anomalias é por meio de técnicas de análise de outlier, usando – por exemplo – gráficos como *Box-plot* ou *Scatter-plot*.** Fora as ferramentas gráficas, podemos utilizar cálculos que podem ser acrescentados às nossas rotinas, tornando o tratamento dos outliers mais eficiente.

A Detecção de Anomalias⁸ [11] é basicamente um valor discrepante, ou seja, um valor que se localiza significativamente distante dos valores considerados normais. É importante notar que uma anomalia não necessariamente é um erro ou um ruído, ela pode caracterizar um valor ou uma classe bem definida, porém de baixa ocorrência, às vezes indesejada, ou que reside fora de agrupamentos ou classes típicas.

Quase todas as bases de dados reais apresentam algum tipo de anomalia, que pode ser causada por fatores como **atividades maliciosas** (por exemplo, furtos, fraudes, intrusões, etc), **erros humanos** (por exemplo, erros de digitação ou de leitura), **mudanças ambientais** (por exemplo, no clima, no comportamento de usuários, nas regras ou nas leis do sistema etc), **falhas em componentes** (por exemplo, peças, motores, sensores, atuadores etc), etc.

As principais aplicações de detecção de anomalias incluem: detecção de fraudes, análise de crédito, detecção de intrusão, monitoramento de atividades, desempenho de rede, diagnóstico de faltas, análise de imagens e vídeos, monitoramento de séries temporais, análise de textos, etc. **A detecção de anomalias em bases de dados é essencialmente um problema de classificação binária, no qual se deseja determinar se um ou mais objetos pertencem à classe normal ou à anômala.**

#VamosPraticar#

(MF – 2011)

Funcionalidade cujo objetivo é encontrar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados. Uma vez encontrados, podem ser tratados ou descartados para utilização em mining. Trata-se de:

- a) descrição.
- b) agrupamento.
- c) visualização.
- d) análise de outliers.
- e) análise de associações.

Comentários: a análise de outliers busca encontrar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados (Letra D).

#VamosPraticar#

Classificação

INCIDÊNCIA EM PROVA: média

Trata-se de uma técnica de mineração de dados que designa itens de dados a uma determinada classe ou categoria previamente definida a fim de prever a classe alvo para cada item de dado.

Galera, a ideia da técnica de classificação é categorizar coisas. Posso dar um exemplo? Existe uma loja americana chamada Target – ela vende de tudo, mas principalmente roupas. Ela ficou famosa em 2002 por conseguir adivinhar mulheres que estavam grávidas e lhes enviar cupons de desconto relacionados a bebês. *Diego, como eles fizeram isso?* Cara, isso virou exemplo de problemas de classificação de livros didáticos de mineração de dados.



A Target precisava classificar cada cliente em uma de duas categorias: provavelmente grávida ou provavelmente não grávida. A classificação é um processo que geralmente funciona em vários estágios. Primeiro, cada instância tem que ser dividida em uma coleção de atributos – também chamados de rótulo ou etiqueta. Para uma loja como a Target, uma instância poderia ser uma mulher qualquer. *Calma que vocês vão entender...*

Eu disse que cada instância deve ser dividida em uma coleção de atributos. *Que atributos seriam relevantes para descobrir se uma mulher está grávida?* Bem, a Target possuía um banco de dados com informações sobre todas as suas clientes como nome, data de nascimento, endereço, e-mail e o principal: histórico de compras. Ademais, é comum a compra ou compartilhamento de bases de dados entre empresas.

Logo, ela possuía todo o histórico de compras de diversas clientes em várias empresas. Além disso, a Target possuía um cadastro em seu site para oferecer descontos para mulheres que se registrassem como grávidas. **Com base em tudo isso, ela analisou uma pequena amostra de dados e, em pouco tempo, surgiram alguns padrões úteis.** Um exemplo interessantíssimo foi a compra de creme hidratante! *Como assim, professor?*

Analisando a base de mulheres grávidas, um analista descobriu que elas estavam comprando quantidades maiores de creme hidratante sem cheiro por volta do começo do segundo trimestre de gravidez. Outro analista observou que, em algum momento das vinte primeiras semanas, mulheres grávidas compraram muitos suplementos como cálcio, magnésio e zinco. **Cada informação dessas pode ser considerada um atributo que ajuda a encaixar nas duas categorias.**

Analistas também descobriram que mulheres na base de registro de grávidas bem próximas de ganhar o bebê estavam comprando de forma diferenciada uma quantidade maior de sabão neutro e desinfetantes para mãos. **Após interpretação, foi descoberto que isso significava que provavelmente o bebê estava próximo de nascer!** *Por que isso é importante?* Porque cada informação dessa é um atributo fundamental para a classificação de gravidez ou não-gravidez.

Claro que, até agora, isso é uma mera expectativa – ainda não é possível dizer que o algoritmo vai acertar na maioria das vezes só porque nós identificamos algumas correlações em uma amostra pequena. Mas, então, foram identificados 25 produtos que – quando analisados em conjunto – permitissem a atribuição de um peso ou uma pontuação. *Como assim, professor?* Basicamente, cada produto (atributo) analisado possui um peso diferente. *Concordam?*

Comprar fraldas é um atributo com um peso/pontuação muito maior do que comprar creme hidratante! Bem, à medida que os computadores começaram a processar os dados, chegou-se a uma pontuação em relação à probabilidade de gravidez de cada compradora! **Mais que isso: foi também possível estimar sua data de nascimento dentro de uma pequena janela, para que a Target pudesse enviar cupons cronometrados para fases mais específicas da gravidez.**

Para testar, criou-se uma pessoa fictícia no banco de dados chamada Jennifer Simpson que tinha 23 anos, morava em Atlanta e em março havia comprado um creme hidratante de cacau, uma bolsa grande o suficiente para caber um pacote de fraldas, suplementos de zinco e magnésio e um tapete azul brilhante. **O algoritmo estimou que havia uma chance de 87% de que ela estivesse grávida e que o bebê nasceria em algum momento no final de agosto.**

Chegou o momento então de colocar o algoritmo para rodar na base histórica inteira! Isso foi feito e a Target começou a enviar cupons de desconto para itens de bebê pelos correios para clientes que nunca haviam se cadastrado como grávidas – tudo baseado na pontuação de seus algoritmos em relação à classificação realizada. **Galera, houve um caso em que um homem irritado entrou em uma Target de Minneapolis exigindo falar com o gerente.**

"Minha filha recebeu isso pelo correio", disse ele. "Ela ainda está no ensino médio e vocês estão enviando cupons para roupas de bebê e berços? Vocês estão tentando incentivá-la a engravidar?". O gerente não tinha ideia do que o homem estava falando. **Ele olhou o cupom e viu que o endereço estava correto, realmente era para a filha daquele senhor e continha anúncios de roupas de maternidade, móveis de criança e fotos de bebês sorridentes.**

O gerente pediu mil desculpas e se despediu do pai da menina. **No dia seguinte, ele – não satisfeito – fez questão de ligar para aquele senhor e pedir desculpas mais uma vez.** Conta-se a história que ao receber o telefone de desculpas, o pai ficou envergonhado: *"Eu tive uma conversa com minha filha"*, disse ele. *"Acontece que tem havido algumas atividades em minha casa que eu não conhecia completamente. Ela está grávida, prevista para agosto e eu que te devo desculpas"*⁹–^[12].



Voltando à classificação! Companhias de Seguro utilizam a classificação para adivinhar quais pacientes idosos morrerão em breve; médicos usam para verificar se bebês prematuros estão desenvolvendo infecções perigosas (já que o classificador pode colocar indicadores sutis de doenças antes que os humanos notem quaisquer sinais); **enfim, há infinitos exemplos de utilização da classificação como técnica de mineração de dados.**

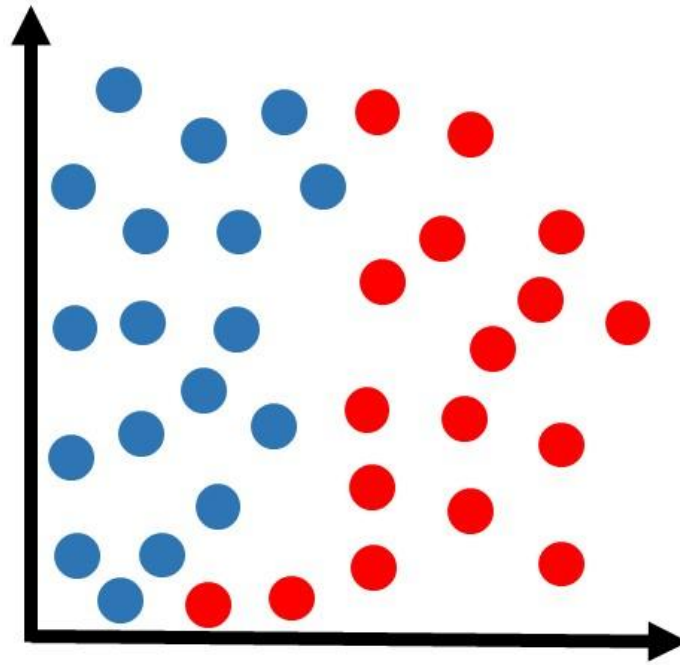
Em suma, podemos dizer que a técnica de classificação utiliza um algoritmo de aprendizado supervisionado a fim de distribuir um conjunto de dados de entrada em categorias ou classes pré-definidas de saída para realizar a análise de dados. **Constroem-se modelos de classificação a partir de um conjunto de dados de entrada, identificando cada classe por meio de múltiplos atributos e os rotulando/etiquetando – sendo essa técnica possível de ser utilizada com outras técnicas!**

Classificador k-NN

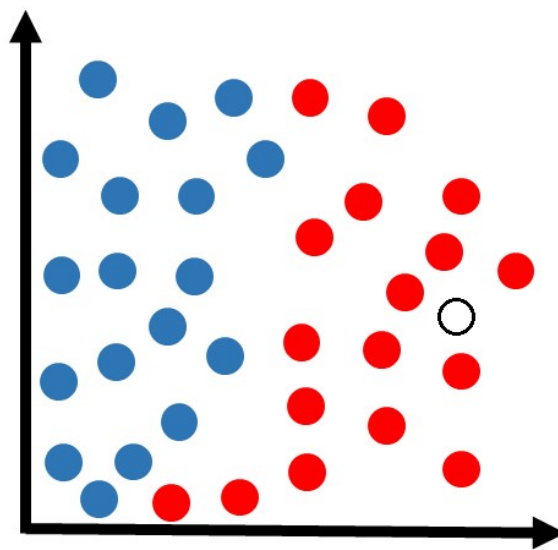
INCIDÊNCIA EM PROVA: baixíssima

Vamos começar falando sobre um dos mais simples algoritmos de classificação: k-NN (k-Nearest Neighbours) – também chamado de k-Vizinhos Mais Próximos. **Trata-se de uma técnica que basicamente separa dados conhecidos em diversas classes a fim de prever a classificação de um novo dado baseado em valores conhecidos mais similares ou mais próximos em termos de distância entre as características.** Calma que ficará claro...

Imagine que recebemos um conjunto de dados (*dataset*) dividido em duas classes distintas: azul e vermelha. Note que já sabemos – de antemão – a classificação das bolinhas dessa amostra:

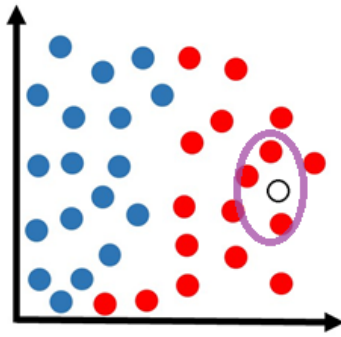


Só que agora eu vou adicionar uma bolinha nova nesse conjunto de dados que eu não sei ainda qual será sua classificação. Aliás, meu objetivo é descobrir qual será a classificação dessa bolinha nova:

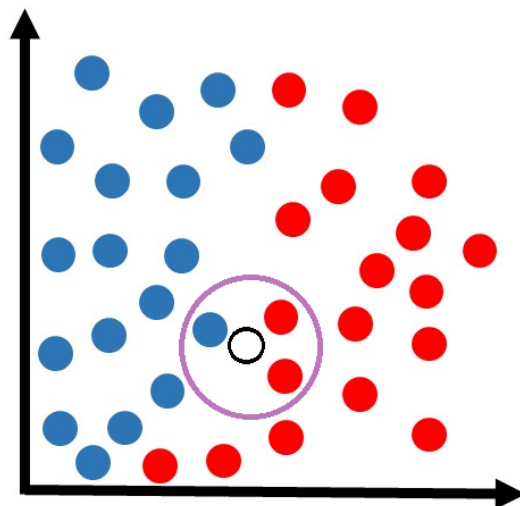
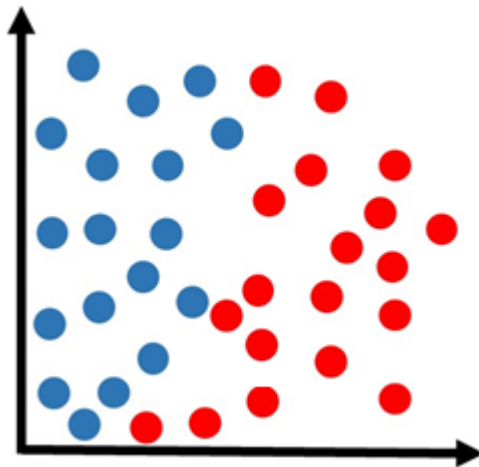


Vocês hão de concordar comigo que – de forma intuitiva – já dá para imaginar que ela tende a ser vermelha. Ora, ela está rodeada de bolinhas vizinhas vermelhas e está mais distante das bolinhas azuis. De toda forma, vamos utilizar o k-NN para fazer essa previsão de forma matemática! *Vocês já sabem que a tradução do nome do algoritmo é k-Vizinhos Mais Próximos, mas o que seria o k?* K é um valor arbitrário escolhido pelo supervisor que indica a quantidade de vizinhos.

Vamos supor que eu seja o supervisor e escolha o valor $k = 2$, logo o algoritmo procurará os 2 vizinhos mais próximos do dado que queremos prever a classificação; se eu escolher o valor $k = 3$, o algoritmo procurará os 3 vizinhos mais próximos do dado que queremos prever a classificação; e assim por diante. No exemplo abaixo, nós procuramos os três vizinhos mais próximos. Vejam só:

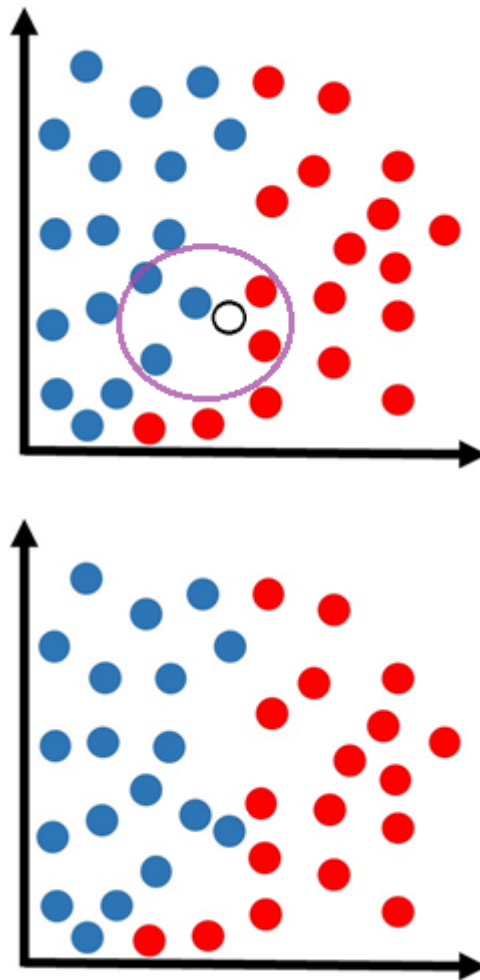


Agora nós faremos uma contagem: nós vamos olhar os três vizinhos mais próximos e vamos contar quantas bolinhas são azuis e quantas bolinhas são vermelhas. Ora, esse exemplo é muito fácil: as três bolinhas ao redor do dado que queremos classificar são vermelhas e nenhuma é azul, então o placar ficou 3×0 , logo o algoritmo vai estimar que o dado que queremos classificar será também uma bolinha vermelha. Agora vamos ver um exemplo um pouquinho mais difícil:



Note que também temos $k = 3$ -Vizinhos Mais Próximos. Se fizermos a contagem das classes dos vizinhos mais próximos, o placar ficará em 2×1 porque temos duas bolinhas vermelhas e uma bolinha azul. Logo, o algoritmo vai estimar que o dado que queremos classificar será

também uma bolinha vermelha. *Simples, né?* Então, vamos complicar um pouquinho mais nosso exemplo alterando o valor de k para $k = 5$:



Opa! Agora o placar mudou: ficou 3x2 para o azul, logo o algoritmo vai estimar que o dado que queremos classificar será também uma bolinha azul. *Você deve estar se perguntando: e se o valor de k for par, não corre o risco de acontecer um empate?* Ocorre, sim! Aí temos algumas opções, quais sejam: a classe da instância mais próxima é atribuída ao novo dado ou simplesmente se atribui uma classe aleatória. *Entendido?*

Agora vamos falar sobre alguns conceitos mais técnicos: k -NN é um classificador de aprendizado supervisionado não paramétrico baseado em distância que pode ser utilizado tanto para classificação quanto para regressão¹⁰_[13]. *Por que é baseado em distância?* Porque o algoritmo se baseia no grau de similaridade entre diversas observações em função de suas características ou variáveis, isto é, calcula a distância de cada um dos pontos em relação ao novo dado que desejamos classificar.

No exemplo, ficou fácil de ver quais são as bolinhas mais próximas, mas o algoritmo utiliza fórmulas de distância euclidiana, distância de cossenos, entre outros para realizar essas medições. *E por que ele é chamado de paramétrico, Diego?* Porque o algoritmo não pressupõe que a relação entre as entradas e as saídas de dados sigam uma função matemática específica, isto é, para cada novo dado que se deseja classificar, utiliza-se o conjunto de dados de treinamento original para calcular.

Complicou um pouco né? Veja só: se eu digo que um conjunto de dados possui uma relação linear entre dados de entrada e saída, eu não preciso armazenar os dados de treinamento para

realizar a previsão de classificação de um novo dado. *Por que?* Porque se o conjunto de dados de treinamento segue uma função específica (Ex: linear), basta eu utilizar a função para realizar a previsão da classificação do novo dado – a isso, chamamos de métodos paramétricos.

Já os métodos não paramétricos não seguem uma função matemática específica. Toda vez que eu quiser realizar a previsão de classificação de um novo dado, eu vou ter que olhar para todos os dados de treinamento – a isso, chamamos de métodos não paramétricos. *O k-NN é um método paramétrico ou não paramétrico?* Como ele não segue nenhuma função matemática específica, ele é considerado um método não paramétrico! Prosseguindo...

Ele também utiliza o que chamamos de aprendizado preguiçoso (*lazy learning*) justamente porque não é necessária uma etapa de treinamento do modelo para somente depois realizar a previsão. Todo o trabalho ocorre no momento em que uma nova previsão é solicitada. É importante falar também sobre o valor de k . Um valor de k pequeno demais gera baixo viés e alta variância (*overfitting*); e um valor k grande demais gera alto viés e baixa variância (*underfitting*).

Ainda não veremos esses conceitos com profundidade. Tudo que você precisa saber por enquanto é que, embora não exista uma regra para se definir k , valores grandes reduzem o efeito dos ruídos na classificação, mas tornam fronteiras de classe menos definidas; já valores pequenos geram modelos mais complexos, o que também traz consequências. Em geral, o valor de k é definido por alguma heurística ou simplesmente por tentativa e erro.

Em outras palavras, o tamanho da vizinhança utilizada para realizar a previsão é arbitrário e definido experimentalmente. Testa-se o modelo com diversos valores de k até encontrar o tamanho ótimo de vizinhança que leva ao menor erro médio. O k-NN é um algoritmo simples com boa acurácia e versátil (pode ser utilizando tanto para classificação quanto para regressão), por outro lado é computacionalmente caro e necessita armazenar os dados de treinamento.

Em suma: o k-NN opera de forma que, dado um objeto x_0 cuja classe se deseja inferir, encontram-se os k objetos x_i , $i = 1, \dots, k$ da base que estejam mais próximos a x_0 e, depois, se classifica o objeto x_0 como pertencente à classe da maioria dos k vizinhos. Logo, ele determina a classe de um objeto com base na classe de outros objetos (instâncias), sendo chamado – portanto – de algoritmo baseado em instância. Vejamos uma questão para solidificar o entendimento...

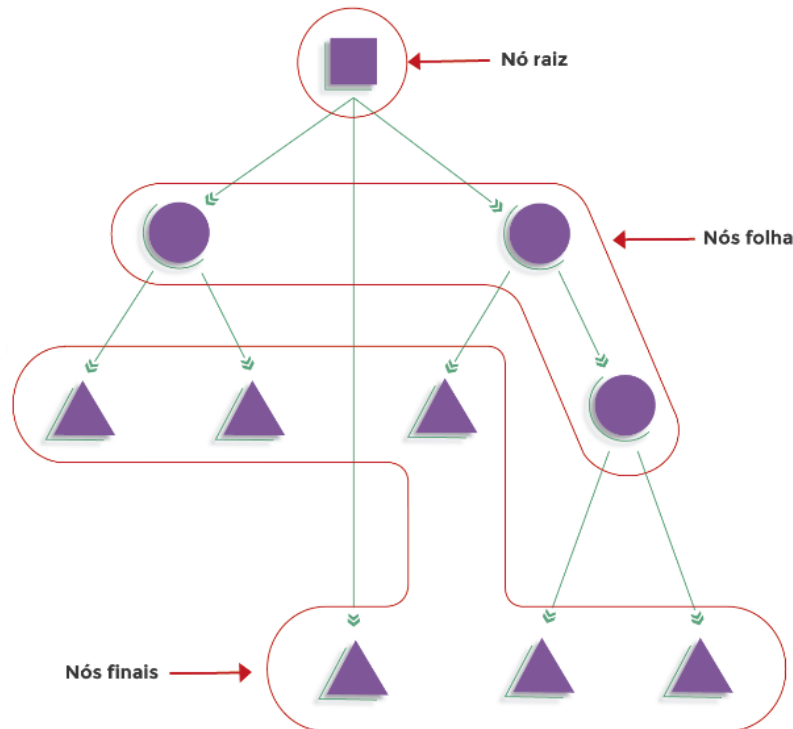
Árvores de Decisão

INCIDÊNCIA EM PROVA: baixíssima

Uma das principais ferramentas de classificação é a árvore de decisão. *O que é isso, Diego?* **É basicamente uma representação gráfica de regras de classificação^[14]!** Elas demonstram visualmente as condições para categorizar dados por meio de uma estrutura que contém nó

raiz, nós folha e nós finais. Na imagem seguinte, eles elementos estão representados respectivamente por um quadrado, um círculo e um triângulo.

É possível atravessar a árvore de decisão partindo do nó raiz até cada folha por meio de diversas regras de decisão – lembrando que o destino final (nó final) contém sempre uma das classes pré-definidas. É importante destacar também que uma árvore de decisão pode ser utilizada tanto para classificação quanto para regressão, mas o nosso foco aqui será na árvore de classificação porque nossas classes são **categóricas e finitas** e, não, **contínuas e infinitas**.



Cada nó interno denota um teste de um atributo, cada ramificação denota o resultado de um teste e cada nó folha apresenta um rótulo de uma classe definida de antemão por um supervisor. O objetivo dessa técnica é criar uma árvore que verifica cada um dos testes até chegar a uma folha, que representa a categoria, classe ou rótulo do item avaliado. *O que isso tem a ver com aprendizado de máquina, Diego?* A árvore de decisão, por si só, não tem a ver com aprendizado de máquina...

No entanto, seu processo de construção automático e recursivo a partir de um conjunto de dados pode ser considerado um algoritmo de aprendizado de máquina. O processo de construção do modelo de uma árvore de decisão se chama indução e busca fazer diversas divisões ou particionamentos dos dados em subconjuntos de forma automática, de modo que os subconjuntos sejam cada vez mais homogêneos. *Como assim, Diego?*

Imagine que eu tenha uma tabela com diversas linhas e colunas, em que as linhas representam pessoas que desejam obter um cartão de crédito e as linhas representam atributos ou variáveis dessas pessoas. Para que uma operadora de cartão decida se vai disponibilizar um cartão de crédito para uma pessoa ou não, ela deve avaliar qual é o risco de tomar um calote dessa pessoa. Logo, a nossa árvore de decisão buscará analisar variáveis para classificar o risco de calote de uma pessoa.

Nós já sabemos que as árvores de decisão são uma das ferramentas do algoritmo de classificação e também sabemos que algoritmos de classificação são supervisionados. Dessa forma, podemos concluir que a árvore de decisão necessita de um supervisor externo para treinar o algoritmo e indicar, de antemão, quais serão as categorias que ele deve classificar uma pessoa. Para o nosso exemplo, vamos assumir que as categorias/classes sejam: Risco Baixo, Risco Médio ou Risco Alto.

Legal! Então, o algoritmo de aprendizado de máquina vai analisar um conjunto de variáveis ou atributos de diversas pessoas e categorizá-las em uma dessas três classes possíveis. *E como o algoritmo vai descobrir quais são as variáveis mais importantes?* De fato, não é uma tarefa simples – ainda mais se existirem muitas variáveis! *A idade é mais relevante para definir o risco de calote de uma pessoa do que seu salário anual?*

O estado civil é mais relevante para definir o risco de calote de uma pessoa do que seu saldo em conta? Em que ordem devemos avaliar cada variável? Para humanos, essas perguntas são extremamente difíceis de responder, mas é aí que entra em cena o aprendizado de máquina! Existe uma lógica interna de construção da árvore de decisão que automaticamente pondera a contribuição de cada uma das variáveis com base em dados históricos. *Como assim, Diego?*

Você pode fazer a máquina aprender oferecendo para ela uma lista que contenha os valores dessas variáveis referentes a diversos clientes antigos e uma coluna extra que indique se esses clientes deram calote na operadora ou não. O algoritmo da árvore de decisão analisará cada uma dessas variáveis (e seus possíveis pontos de corte) a fim de descobrir quais são as melhores para realizar o particionamento dos dados de modo que se formem dois subgrupos mais homogêneos possíveis.

O que você quer dizer com grupos mais homogêneos possíveis, professor? Galera, esse é o momento em que o algoritmo fará diversos testes: ele poderá começar analisando, por exemplo, a variável de estado civil. Ele separa as pessoas em dois subgrupos (casados e não-casados) e verifica qual é a porcentagem de casados caloteiros e não-casados caloteiros. Pronto, ele anota esse valor para poder fazer diversas comparações posteriormente.

Depois ele pode analisar a idade: separa em dois subgrupos (< 25 anos e ≥ 25 anos) e verifica qual é a porcentagem de < 25 anos caloteiros e ≥ 25 anos caloteiros. Pronto, ele anota esse valor também. *Professor, por que 25 anos?* Eu dei apenas um chute, mas o algoritmo pode fazer diversos testes e verificar o melhor corte. *Vocês entenderam a lógica?* Pois é, o algoritmo fará isso para todas as variáveis!

Agora vem o pulo do gato: eu disse para vocês que os subgrupos formados devem ser os mais homogêneos possíveis. Quando ele dividiu a primeira variável em casados caloteiros e não-casados caloteiros, ele pode ter descoberto que havia muito mais não-casados caloteiros do que casados caloteiros. Logo, o fato de uma pessoa ser casada tende a reduzir seu risco de calote. Pessoal, é claro que isso aqui é uma simplificação...

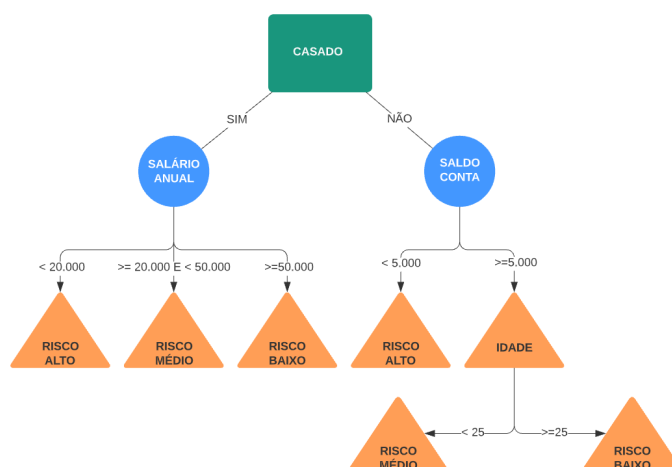
O algoritmo vai fazer milhares de testes com cada uma das variáveis, vai testar diversos pontos de corte diferentes e diversas sequências de análise de variáveis diferentes. O que importa aqui é que nós vamos sair de um grupo muito misturado (menos homogêneo) para dois subgrupos

menos misturados (mais homogêneos). *Sabe qual é o nome disso no contexto de ciência de dados?* Ganho de Informação ou Redução de Entropia.

Quem sabe o que é entropia? Entropia é a uma medida que nos diz o quanto um conjunto de dados está desorganizado ou misturado. Ora, toda vez que nós particionamos os dados em subgrupos, nós obtemos dados mais homogêneos e organizados, logo nós reduzimos a entropia. O que a nossa árvore de decisão busca fazer é pegar um conjunto de dados e encontrar um conjunto de regras sobre variáveis ou pontos de corte que permite separar esses dados em grupos mais homogêneos.

Note que sempre que dividimos um grupo em outros subgrupos, esses subgrupos serão mais puros à medida que seus dados forem mais homogêneos. Em outras palavras, quanto mais homogêneos forem os dados de um subconjunto, mais puros serão seus dados. Para calcular a pureza de um subgrupo, isto é, quão homogêneos são seus dados, existem diversas métricas (Ex: Índice de Gini, Redução de Variância, etc).

Ao final do treinamento da máquina, chegamos a um possível modelo de árvore de decisão efetivamente construído e disponível para ser utilizado com novos clientes. Vejamos:



O algoritmo inicialmente faz uma avaliação se a pessoa é solteira ou casada. Caso ela casada: se seu salário anual for abaixo de 20.000, é classificada como risco alto; se seu salário anual for entre 20.000 e 50.000, é classificada como risco médio; se seu salário anual for maior que 50.000, é classificada como risco baixo. Caso não seja casada: se seu saldo em conta for abaixo de 5.000, é classificada como risco alto; se seu saldo em conta for acima de 5.000, analisa-se sua idade.

Se ela não é casada, possui mais de 5.000 em conta, mas tem menos de 25 anos, é classificada como risco médio; se ela não é casada, possui mais de 5.000 em conta, mas tem mais de 25 anos, é classificada como risco baixo. *Gostaram da nossa árvore?* Acho que ela faz sentido! Parece ser um bom modelo para avaliar risco de calote e decidir se deve receber um cartão de crédito ou não. *Ela é perfeita?* Não, nenhum modelo será perfeito!

Eu disse páginas atrás que o processo de construção do modelo de uma árvore de decisão busca fazer diversas divisões ou particionamentos dos dados em subconjuntos de forma automática, de modo que os subconjuntos sejam cada vez mais homogêneos. Ora, imaginem

um contexto em que tenhamos dezenas de variáveis! Se o objetivo do algoritmo é obter automaticamente subconjuntos cada vez mais homogêneos, então temos um problema grave...

Quando o algoritmo vai parar de fazer as divisões? Em tese, ele pode ir dividindo, dividindo, dividindo indefinidamente até que – no pior caso – tenhamos um único dado para cada nó folha. O nome desse fenômeno é *overfitting*! Ele deve ser evitado porque pode tornar o modelo de árvore de decisão completamente inútil. *E o que devemos fazer, professor?* É necessário estabelecer um limite para as divisões...

Há diversas maneiras: nós podemos definir uma altura/profundidade máxima da árvore – quando esse limite for atingido, interrompe as subdivisões; nós também podemos realizar a poda da árvore – deixamos a árvore crescer quanto quiser e depois vamos reduzindo as divisões que sejam pouco significativas (é como se realmente fizéssemos uma poda de uma árvore). Por fim, vamos ver algumas vantagens e desvantagens:

VANTAGENS DE ÁRVORES DE DECISÃO

As árvores de decisão podem gerar regras compreensíveis e executam a classificação sem exigir muitos cálculos, sendo capazes de lidar com variáveis contínuas e categóricas.

As árvores de decisão fornecem uma indicação clara de quais campos são mais importantes para predição ou classificação.

As árvores de decisão são menos apropriadas para tarefas de estimativa em que o objetivo é prever o valor de um atributo contínuo.

As árvores de decisão estão sujeitas a erros em problemas de classificação com muitas classes e um número relativamente pequeno de exemplos de treinamento.

Por meio da técnica de estratificação (estrato = camadas), é capaz designar regras para cada caso a uma dentre várias categorias diferentes.

DESVANTAGENS DE ÁRVORES DE DECISÃO

As árvores de decisão são bastante propensas ao overfitting dos dados de treinamento.

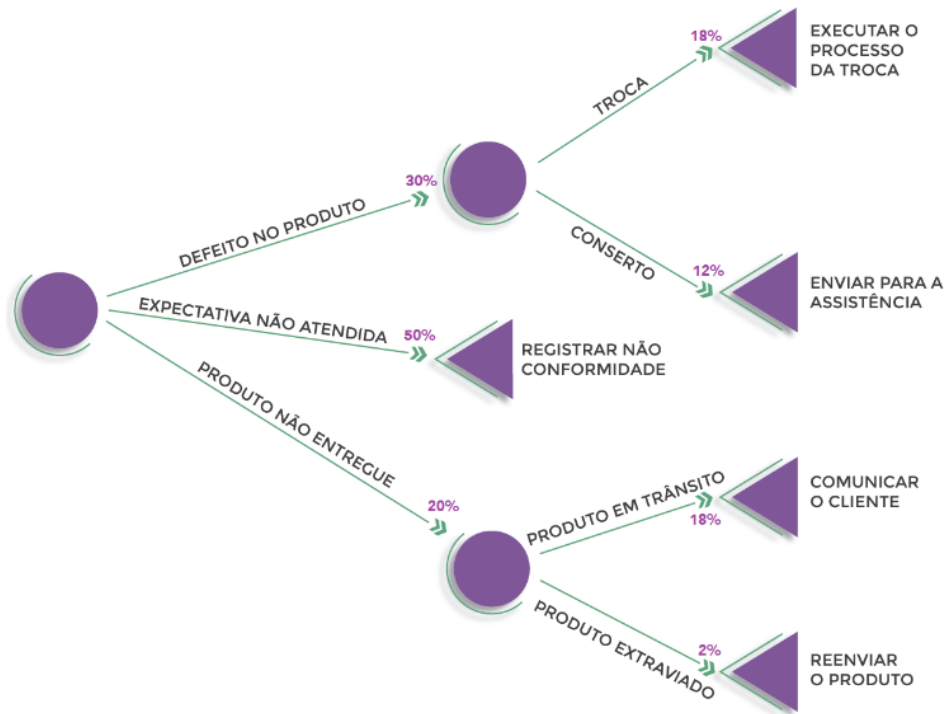
Uma única árvore de decisão normalmente não faz grandes previsões, portanto várias árvores são frequentemente combinadas em forma de florestas chamadas (Random Forests).

Somente se as informações forem precisas e exatas, a árvore de decisão fornecerá resultados promissores. Mesmo se houver uma pequena alteração nos dados de entrada, isso pode

causar grandes alterações na árvore.

Se o conjunto de dados é enorme, com muitas colunas e linhas, é uma tarefa muito complexa projetar uma árvore de decisão com muitos ramos.

Se uma das regras do modelo estiver incorreta, isso gerará divisões equivocadas da árvore, fazendo com que o erro se propague por todo o resto da árvore.



É importante mencionar que esse algoritmo é capaz de classificar dados dentre de um conjunto finito de classes com base em valores de entrada por meio de uma abordagem chamada **estratificação**, que permite determinar as regras para que se possa designar ou direcionar cada caso a uma categoria pré-existente, separando-os em níveis diferentes (Ex: executar o processo da troca, enviar para a assistência, comunicar o cliente, reenviar o produto). *Bacana?*

Os algoritmos de árvore de decisão mais comuns são: ID3, c4.5 e CART. O algoritmo ID3 é utilizado para gerar árvores a partir de um conjunto de dados. É usado em problemas de aprendizado supervisionado, onde os dados de entrada são divididos em categorias com base em certas condições. O algoritmo funciona selecionando o melhor atributo de um determinado conjunto de atributos para dividir os dados em dois (ou mais) subconjuntos.

Isso é feito de forma iterativa, onde – a cada passo do algoritmo – o melhor atributo é escolhido e os dados são divididos em dois (ou mais) subconjuntos. Este processo é repetido até que todos os dados sejam divididos em categorias. Depois que a árvore é gerada, ela pode ser usada para classificar novos pontos de dados seguindo o caminho da árvore até que o ponto de dados seja associado a uma categoria.

Ela trabalha apenas com atributos categóricos/qualitativos (Ex: Masculino ou Feminino), construindo a árvore a partir do nó raiz de cima para baixo recursivamente por meio do método dividir-para-conquistar. Já o algoritmo c4.5 é uma evolução do ID3 capaz de trabalhar tanto com atributos categóricos quanto numéricos/quantitativos (Ex: 95kg). Ele funciona dividindo iterativamente os dados em subconjuntos com base no atributo que melhor separa os dados.

A cada divisão, o C4.5 calcula a impureza de cada um dos subconjuntos e seleciona o atributo que produz os subconjuntos mais puros, isto é, aqueles que contêm apenas dados de apenas uma classe ou rótulo. O algoritmo então repete o processo até que todos os dados tenham sido divididos em subconjuntos puros ou até que um critério de parada seja atingido. Além disso, trata-se de um algoritmo bem mais rápido que o anterior.

Por fim, o algoritmo CART (*Classification and Regression Trees*) é um tipo de algoritmo de árvore de decisão utilizado para problemas de classificação e regressão. Ele funciona construindo uma árvore de decisão durante a fase de treinamento, que é usada para fazer previsões sobre dados não vistos. O objetivo do algoritmo é criar um modelo que preveja com precisão o valor alvo, além de ter o menor número possível de divisões.

Ele funciona dividindo recursivamente os dados de treinamento em subconjuntos menores com base na variável independente mais significativa. A árvore para de crescer quando atinge uma profundidade máxima especificada ou quando todas as amostras restantes pertencem à mesma classe. O algoritmo CART é capaz de trabalhar tanto com atributos categóricos/qualitativos quanto numéricos/quantitativos.

Florestas Aleatórias

RELEVÂNCIA EM PROVA: baixíssima

Árvores de decisão são ferramentas interessantes, mas não fazem previsões com acurácia. Em outras palavras, elas funcionam bem com dados de treinamento, mas não são tão flexíveis quando utilizadas para classificar novas amostras. Já florestas aleatórias (*random forests*) permitem combinar a simplicidade de árvores de decisão com flexibilidade para melhorar significativamente a acurácia das previsões, sendo utilizadas tanto para classificação quanto para regressão.

Dor no	BOA	ARTÉRIAS	PESO	DOENÇA
Peito	CIRCULAÇÃO	BLOQUEADAS		CARDÍACA

Não	Não	Não	125	Não
Sim	Sim	Sim	180	Sim
Sim	Sim	Não	210	Não
Sim	Não	Sim	167	Sim

Vamos ver como isso funciona? Imagine que nós tenhamos um conjunto de dados (*dataset*) original com diversas variáveis mostrado acima. Porém, em nosso primeiro passo, precisaremos criar um *bootstrapped dataset*. O que diabos é isso, Diego? É um conjunto de dados retirados do conjunto de dados original (inclusive com o mesmo tamanho do conjunto de dados original), mas com amostras aleatórias dos dados. Aqui já temos um ponto de atenção...

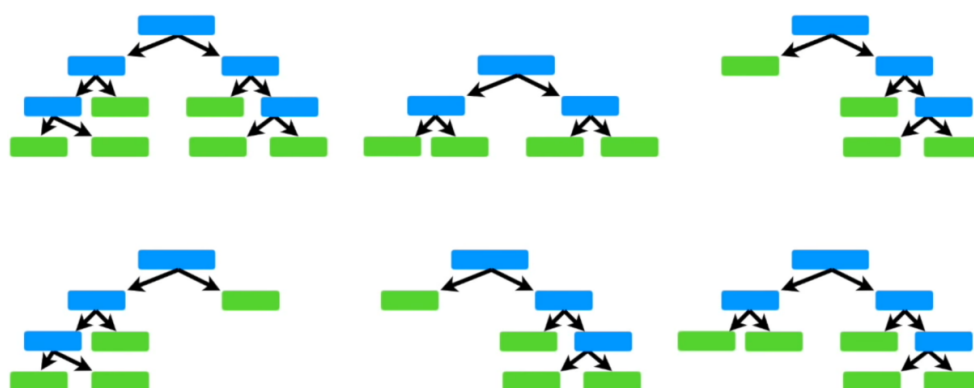
Você deve estar imaginando: se o *bootstrapped dataset* é um conjunto de dados **retirado** do conjunto de dados original e tem o **mesmotamanho**, então deve ser uma cópia dos dados originais. Não, porque as amostras de dados retirada do conjunto de dados original pode conter amostras repetidas. Vejam um exemplo na tabela a seguir: a primeira linha é igual a segunda linha do *dataset* original; a segunda linha é igual a primeira do *dataset* original.

dataset original				
Dor no Peito	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA
Não	Não	Não	125	Não
Sim	Sim	Sim	180	Sim
Sim	Sim	Não	210	Não
Sim	Não	Sim	167	Sim
Bootstrapped dataset				
Dor no Peito	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA

Sim	Sim	Sim	180	Sim
Não	Não	Não	125	Não
Sim	Não	Sim	167	Sim
Sim	Não	Sim	167	Sim

Note que a quarta linha do conjunto de dados original foi repetida duas vezes e a terceira sequer foi selecionado. É por essa razão que o *bootstrapped dataset* é considerado um *dataset* retirado do *dataset* original e possui o mesmo tamanho. O segundo passo é criar uma árvore de decisão baseado no *bootstrap dataset*, mas com uma particularidade: nós vamos utilizar um subconjunto aleatório de colunas para cada nível da árvore.

Por exemplo: para a primeira árvore, nós vamos escolher duas variáveis quaisquer e verificar qual delas divide melhor as amostras. Vamos supor que tenhamos escolhido “Boa Circulação” e “Artérias Bloqueadas” e que, dentre essas duas, “Boa Circulação” divide melhor as amostras. Agora vamos para o próximo nível e fazemos os mesmos passos até utilizar todas as variáveis disponíveis. Essa iteração permitirá construir diversas árvores diferentes:



É a variedade que torna as florestas aleatórias mais efetivas que árvores de decisão utilizadas de forma individual. Bem, agora que nós temos uma enorme variedade de árvores de decisões (uma enorme variedade de árvores é uma... floresta), nós podemos tentar prever de uma determinada pessoa tem ou não doenças cardíacas. Considere abaixo os dados de uma nova pessoa que deseja prever doenças cardíacas:

Novo paciente				
Dor no Peito	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA
Sim	Não	Não	168	?

Nós pegamos esses dados das variáveis e passamos por todas as árvores de decisão aleatórias criadas nos passos anteriores. Vamos supor que, de 100 árvores aleatórias, 75 indicaram que esse paciente possui doença cardíaca e 25 indicaram que ele não possui. Dessa forma, podemos prever que – sim – ele tem doença cardíaca. *Agora, como nós sabemos que a floresta aleatória fez um bom trabalho de previsão de dados?* Essa é a parte legal...

dataset original				
Dor no Peito	BOA CIRCULAÇÃO	ARTÉRIAS BLOQUEADAS	PESO	DOENÇA CARDÍACA
Não	Não	Não	125	Não
Sim	Sim	Sim	180	Sim
Sim	Sim	Não	210	Não
Sim	Não	Sim	167	Sim

Vocês se lembram que no início da explicação eu disse que o *bootstrapped dataset* tinha o mesmo tamanho do *dataset original* e permitia dados duplicados? A consequência disso é que algumas amostras do *dataset original* não foram utilizadas (em geral, 1/3 das amostras de *datasets* originais não são utilizadas no *bootstrapped dataset*). Em nosso exemplo, a 3ª linha não foi utilizada, logo eu posso usar os dados reais das variáveis, passar pela floresta aleatória e ver se os resultados batem.

Assim, para saber a acurácia da previsão de dados, basta rodar os dados das variáveis da terceira linha e verificar se a floresta também retorna: **Não**. *Fechado?* Em relação às árvores de decisão individuais, podemos afirmar que esse algoritmo é capaz de fazer previsões mais acuradas, dado que ele combina previsões de diversas árvores de decisão treinadas em diferentes subconjuntos do conjunto de dados de treinamento original.

Ademais, ela é menos propensa a sofrer com *overfitting*, dado que árvores individuais geralmente se ajustam excessivamente aos dados de treinamento. Por fim, alguns conceitos importantes:

Características das florestas de dados

A técnica que combina múltiplos *bootstrapping* de dados com uma agregação final é chamada de *bagging* (*bootstrapping aggregation*).

Florestas aleatórias são capazes de resolver problemas de classificação ou regressão e funcionam bem com variáveis categóricas ou contínuas.

A utilização de florestas aleatórias permite reduzir o *overfitting* (veremos mais à frente) ao reduzir a variância e, portanto, melhorar a acurácia;

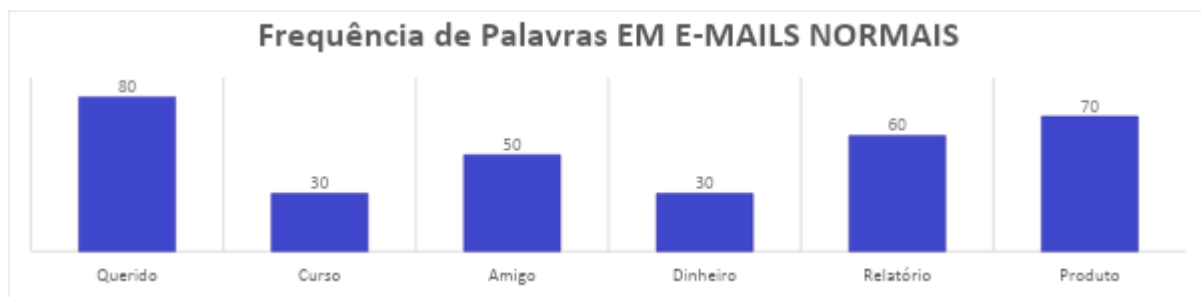
Trata-se de um algoritmo bastante robusto a ruídos e dados anômalos, além de conseguir lidar bem com grandes conjuntos de dados.

Classificador Naive Bayes

RELEVÂNCIA EM PROVA: baixíssima

Para explicar o que é um Classificador Naive Bayes, eu preciso fazer uma contextualização! Deem uma olhada na caixa de entrada de seus e-mails agora. Em geral, há e-mails normais e e-mails de spam. Os e-mails normais são aquelas de amigos, colegas de trabalho e familiares; já os e-mails indesejados são aqueles de spam com tentativas de golpes ou anúncios não solicitados. *O que deveríamos fazer se quiséssemos filtrar as mensagens de spam?*

Nós podemos pegar o texto de todos os e-mails normais (aquelas de amigos, colegas e familiares) e criar um gráfico com a frequência de todas as palavras em suas mensagens. Exemplo...



Em seguida, nós podemos utilizar esse gráfico para calcular as probabilidades de encontrar cada palavra dado que elas estão presentes em um e-mail normal. Por exemplo: a probabilidade de vermos a palavra “Querido” dado que estamos vendo apenas e-mails normais é $80/320$. *Por que?* Porque essa probabilidade é calculada pela frequência da palavra “querido” sobre a quantidade total de palavras contidas em todos os e-mails normais¹²_[15] da seguinte forma:

$$P(\text{Normais}) = \frac{\text{Frequência(Querido)}}{\text{Total de Palavras}} = \frac{80}{(80+30+50+30+60+70)} = \frac{80}{320} = 0,25$$

Se fizermos esse mesmo procedimento para as outras palavras contidas nos e-mails normais, vamos obter os seguintes resultados:

$$P(\text{Normais}) = \frac{\text{Frequência(Curso)}}{\text{Total de Palavras}} = \frac{30}{(80+30+50+30+60+70)} = \frac{30}{320} = 0,09$$

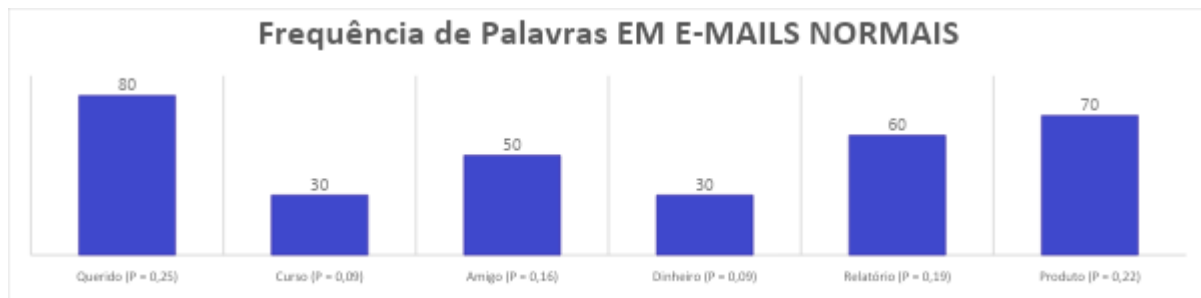
$$P(\text{Normais}) = \frac{\text{Frequência(Amigo)}}{\text{Total de Palavras}} = \frac{50}{(80+30+50+30+60+70)} = \frac{50}{320} = 0,16$$

$$P(\text{Normais}) = \frac{\text{Frequência(Dinheiro)}}{\text{Total de Palavras}} = \frac{30}{(80+30+50+30+60+70)} = \frac{30}{320} = 0,09$$

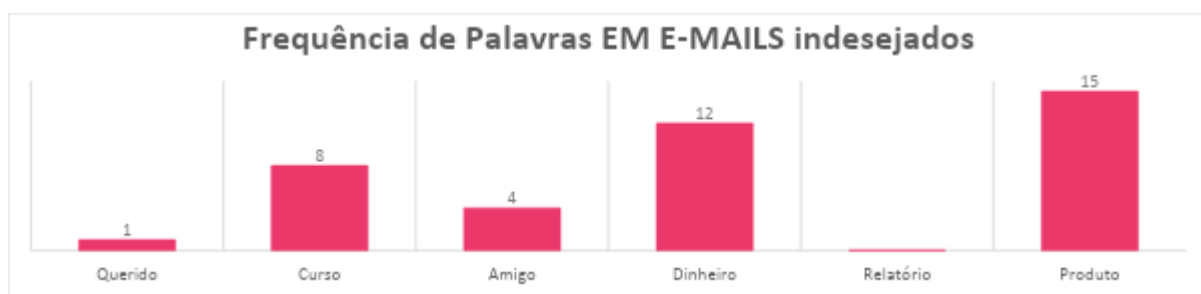
$$P(\text{Normais}) = \frac{\text{Frequência(Relatório)}}{\text{Total de Palavras}} = \frac{60}{(80+30+50+30+60+70)} = \frac{60}{320} = 0,19$$

$$P(\text{Normais}) = \frac{\text{Frequência(Produto)}}{\text{Total de Palavras}} = \frac{70}{(80+30+50+30+60+70)} = \frac{70}{320} = 0,22$$

Pronto! Agora para que nós não esqueçamos qual é a probabilidade de cada palavra, vamos colocá-las no gráfico de frequência de palavras em e-mails normais:



Agora vamos fazer a mesma coisa, mas para e-mails de spam! Para tal, pegamos o texto de todos os e-mails de spam e criamos um gráfico com a frequência de todas as palavras em suas mensagens:



Da mesma forma, nós podemos calcular a probabilidade de cada uma das palavras, porém agora dentro do contexto de e-mails indesejados¹³ [16]:

$$P(\text{Indesejadas}) = \frac{\text{Frequência(Querido)}}{\text{Total de Palavras}} = \frac{1}{(1+8+4+12+0+15)} = \frac{1}{40} = 0,03$$

Se fizermos esse mesmo procedimento para as outras palavras contidas nos e-mails indesejados, vamos obter os seguintes resultados:

$$P(\text{Indesejadas}) = \frac{\text{Frequência(Curso)}}{\text{Total de Palavras}} = \frac{8}{(1+8+4+12+0+15)} = \frac{8}{40} = 0,20$$

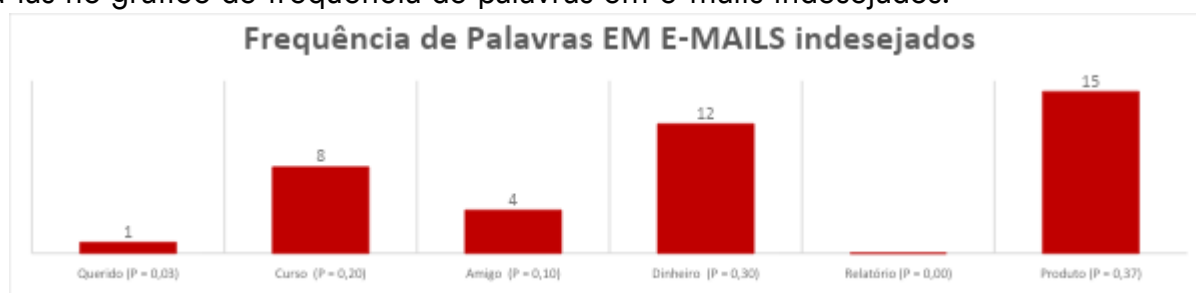
$$P(\text{Indesejadas}) = \frac{\text{Frequência(Amigo)}}{\text{Total de Palavras}} = \frac{4}{(1+8+4+12+0+15)} = \frac{4}{40} = 0,10$$

$$P(\text{Indesejadas}) = \frac{\text{Frequência(Dinheiro)}}{\text{Total de Palavras}} = \frac{12}{(1+8+4+12+0+15)} = \frac{12}{40} = 0,30$$

$$P(\text{Indesejadas}) = \frac{\text{Frequência(Relatório)}}{\text{Total de Palavras}} = \frac{0}{(1+8+4+12+0+15)} = \frac{0}{40} = 0,00$$

$$P(\text{Indesejadas}) = \frac{\text{Frequência(Produto)}}{\text{Total de Palavras}} = \frac{15}{(1+8+4+12+0+15)} = \frac{15}{40} = 0,37$$

Pronto! Agora para que nós não esqueçamos qual é a probabilidade de cada palavra, vamos colocá-las no gráfico de frequência de palavras em e-mails indesejados:



PROBABILIDADES DE E-MAILS NORMAIS	PROBABILIDADES DE E-MAILS INDESEJADOS
 	

			
$P(\text{Querido} \text{Normal})$	0,25	$P(\text{Querido} \text{indesejado})$	0,03
$P(\text{curso} \text{Normal})$	0,09	$P(\text{curso} \text{indesejado})$	0,20
$P(\text{amigo} \text{Normal})$	0,16	$P(\text{amigo} \text{indesejado})$	0,10
$P(\text{dinheiro} \text{Normal})$	0,09	$P(\text{dinheiro} \text{indesejado})$	0,30
$P(\text{relatório} \text{normal})$	0,19	$P(\text{relatório} \text{indesejado})$	0,00
$P(\text{Produto} \text{normal})$	0,22	$P(\text{Produto} \text{indesejado})$	0,37

Bacana! Agora imagine que eu acabo de receber uma mensagem que começa com:

“Querido amigo,

(...)”

Como saber se esse e-mail é normal ou indesejado? Vamos começar dando um palpite inicial sobre a probabilidade de um e-mail qualquer (independentemente do que diz em seu texto) seja um e-mail normal. Como é um palpite, podemos sugerir qualquer probabilidade que quisermos (Ex: 50%), mas um palpite mais direcionado pode ser estimado a partir do nosso conjunto de dados de treinamento (Ex: 89%). Como assim, Diego?

Vamos lá! Estamos querendo descobrir a probabilidade de um determinado e-mail ser um e-mail normal. Logo, podemos chutar – por exemplo – que existe uma probabilidade de 50% desse e-mail ser um e-mail normal. Por outro lado, nós podemos dar um chute mais direcionado. Ora, se nosso conjunto de treinamento continha 320 e-mails normais e 40 e-mails indesejados, logo um palpite inicial mais razoável seria:

$$P_{\text{Normal}} = \frac{\text{Quantidade (Normal)}}{\text{Quantidade (Total)}}$$

$$P_{\text{Normal}} = \frac{320}{360} = 0,89$$

Como queremos calcular a probabilidade de um e-mail que contenha as palavras “Querido amigo”, devemos multiplicar nosso palpite inicial pela probabilidade de a palavra “Querido” ocorrer em um e-mail normal e multiplicar pela probabilidade de a palavra “amigo” ocorrer em um e-mail normal. Ora, nós temos essas probabilidades todas calculadas da etapa anterior em nossa tabelinha apresentada na página passada. Vamos fazer os cálculos...

$$P_{\text{Normal}} \times P_{\text{Normal}} \times P(\text{Amigo}|\text{Normal}) = 0,89 \times 0,25 \times 0,16 = 0,0356 = 3,56\%$$

Agora vamos repetir o procedimento para e-mails indesejados. Começamos pelo cálculo do nosso palpite inicial:

$$P_{\text{Indesejado}} = \frac{\text{Quantidade (Indesejado)}}{\text{Quantidade (Total)}}$$

$$P_{\text{Indesejado}} = \frac{40}{360} = 0,11$$

Como queremos calcular a probabilidade de um e-mail que contenha as palavras “Querido amigo”, devemos multiplicar nosso palpite inicial pela probabilidade de a palavra “Querido” ocorrer em um e-mail indesejado e multiplicar pela probabilidade de a palavra “amigo” ocorrer em um e-mail indesejado. Ora, nós temos essas probabilidades todas calculadas da etapa anterior em nossa tabelinha apresentada na página passada. Vamos fazer os cálculos...

$$P_{\text{Indesejado}} \times P_{\text{Indesejado}} \times P(\text{Amigo}|\text{Indesejado}) = 0,11 \times 0,03 \times 0,1 = 0,00033 = 0,03\%$$

Ora, como a probabilidade que calculamos para que essa mensagem fosse um e-mail normal é maior que a probabilidade que calculamos para que essa mensagem fosse um e-mail indesejado (**3,56%** > **0,03%**), logo esse e-mail será considerado normal e, não, indesejado. Pronto, vocês acabaram de entender o funcionamento prático desse classificador. Agora vamos ver algumas formalizações teóricas...

O Classificador Naive Bayes é um classificador probabilístico baseado no Teorema de Bayes com hipótese forte de independência entre seus atributos/variáveis. *O que é o Teorema de Bayes?* Trata-se de um teorema que descreve a probabilidade condicional de um evento, baseado em um conhecimento anterior que pode estar relacionado ao evento. A fórmula desse teorema é extremamente simples, sendo basicamente uma aplicação direta de sua definição:

$$P(A|B) = \frac{P(A) \times P(B)}{P(A)}$$

Esse teorema pode ser lido como: probabilidade de ocorrência de um evento A dado que um evento B ocorreu é igual à probabilidade de ocorrência de um evento B dado que um evento A ocorreu multiplicado pela probabilidade de ocorrência de um evento A sobre a probabilidade de ocorrência de um evento B. De forma mais simples, a probabilidade de A sabendo B é igual à probabilidade de B sabendo A, multiplicado pela probabilidade de A e dividido pela probabilidade de B.

Parece complicado, mas não é! Esse teorema permite para calcular uma probabilidade condicional, isto é, probabilidade de ocorrência de um evento dado que outro evento ocorreu. Por exemplo: *qual é a probabilidade de uma pessoa qualquer ser uma mulher dado que essa pessoa tem 1,90m de altura?* Pelo Teorema de Bayes, é a probabilidade de uma pessoa ter 1,90m de altura dado que ela é uma mulher, multiplicado pela probabilidade de ser mulher dividido pela probabilidade de ter 1,90m.

O Teorema de Bayes é a base para estudar o Classificador Naive Bayes. A palavra de origem inglesa *naive* significa ingênuo, ou seja, esse tópico trata do classificador ingênuo de Bayes. *Por que ele é chamado de ingênuo?* Porque pressupõe-se que as variáveis ou atributos contribuem para a probabilidade de forma independente uma da outra. Em outras palavras, trata-se de uma aplicação ingênua do Teorema de Bayes por considerar que as variáveis são independentes.

Na prática, isso significa que esse classificador retorna a mesma probabilidade independente da ordem dos eventos. Logo, a probabilidade para “Querido Amigo” é igual à de “Amigo Querido”:

$$P(\text{Normal}) \times P(\text{Normal}) \times P(\text{Amigo}|\text{Normal}) = 0,89 \times 0,25 \times 0,16 = 0,0356 = \mathbf{3,56\%}$$

$$P(\text{Normal}) \times P(\text{Normal}) \times P(\text{Querido}|\text{Normal}) = 0,89 \times 0,16 \times 0,25 = 0,0356 = \mathbf{3,56\%}$$

Nós temos regras gramaticais que regulam a forma em que falamos ou escrevemos, mas esse classificador ignora completamente essa característica. Ele considera que a língua escrita é simplesmente um saco cheio de palavras e cada mensagem de e-mail é só um conjunto aleatório dessas palavras, mas – de forma curiosa – esse classificador funciona bem com aprendizado de máquina. Logo, diz-se que ele tem alto viés e baixa variância (veremos em outro tópico).

Note que existe um paralelismo interessante entre esse classificador e o algoritmo de árvores de decisão, visto que ambos respondem perguntas para descobrir a probabilidade de algo pertencer a uma determinada classe (Ex: um e-mail com as palavras “querido” e “amigo” ser um spam). Só que, para o algoritmo de árvore de decisão, a ordem das perguntas importa; no classificador ingênuo, a ordem das perguntas não importa porque as variáveis são consideradas não correlacionáveis.

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Por fim, um dos grandes benefícios desse classificador é que ele é muito simples, rápido e escalável, dado que as variáveis são independentes (e isso facilita muito os cálculos!). Em alguns casos, a velocidade é preferível à maior precisão. Além disso, ele funciona bem com a classificação de texto, processamento de linguagem natural, detecção de spam, entre outros. Ele também é capaz de realizar, com precisão, o treinamento de um modelo com uma quantidade reduzida de amostras.

Como desvantagens, podemos destacar que ele assume que as variáveis são independentes, o que raramente ocorre na vida real e há o problema de variáveis com nenhuma ocorrência de frequência. Em suma: o Classificador Naive Bayes se baseia na probabilidade condicional do Teorema de Bayes para encontrar a mais provável das possíveis classificações considerando que as variáveis/atributos não são correlacionadas, isto é, são independentes entre si (ou ingênuos).

Support Vector Machine (SVM)

RELEVÂNCIA EM PROVA: baixíssima

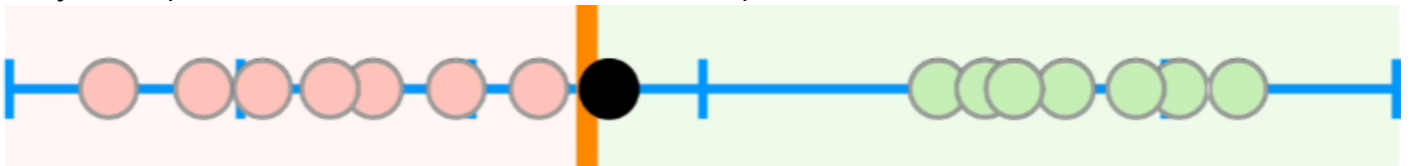
SVM (*Support Vector Machine*) é método de mineração de dados de aprendizado supervisionado não probabilístico utilizado tanto para classificação quanto para regressão. Para entender seu funcionamento, vamos partir de um exemplo de peso de ratos de laboratório. Suponha que colemos o peso de cada rato e coloquemos seus valores em uma linha de tal forma que as bolinhas vermelhas representem ratos não obesos e as bolinhas verdes representem ratos obesos.



Dessa maneira, nós podemos escolher um valor limite de tal forma que, caso uma nova observação possua um peso menor que esse valor limite, ela será classificada como não obesa; e caso possua um peso maior, será classificada como obesa. Esse valor limite é representado na imagem seguinte pela barra laranja (apesar de ser um ponto). Note que qualquer nova observação à esquerda da barra alaranjada será considerada não obesa e qualquer observação à direita será considerada obesa.



Infelizmente as coisas não são tão simples assim! *E se tivermos uma nova observação que esteja bem próximo do valor limite como a bolinha preta abaixo?*



Ora, está à direita do valor limite, logo deveria ser classificada com obesa. No entanto, isso não faz muito sentido porque ela está mais próxima dos não obesos do que dos obesos. Podemos concluir que esse valor limite que escolhemos não foi muito legal! *Vamos escolher outro?* Uma maneira interessante seria focar nas observações que estão localizadas na borda interna de cada classe de dados conforme apresenta a imagem e utilizar o ponto central entre as duas como limite:

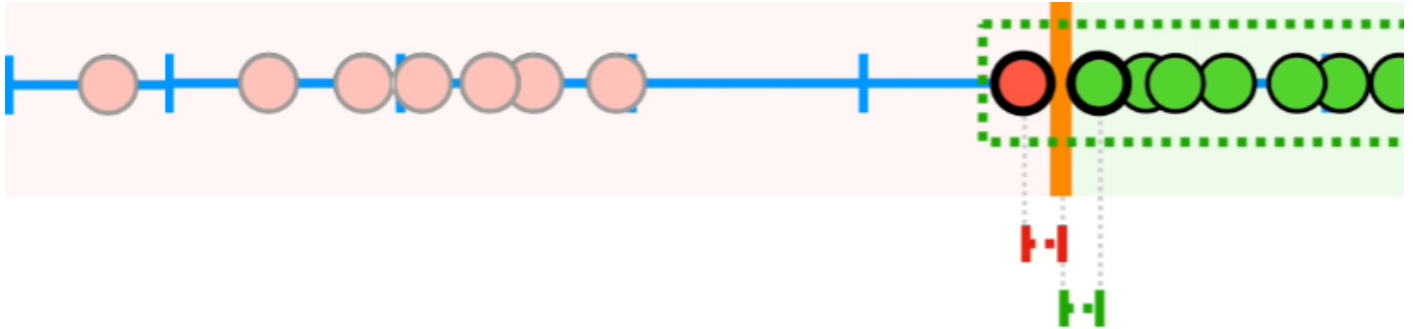


Note que agora, se uma nova observação estiver localizada à esquerda do valor limite, ela estará sempre mais próxima das observações classificadas como não obesas do que das observações classificadas como obesas conforme apresenta a imagem seguinte, logo faz todo sentido classificá-la realmente como não obesa. E é claro que isso ocorre também de forma análoga para as observações classificadas como obesas:



A menor distância entre observações e o valor limite é chamada de margem – isso será importante daqui a pouco. Quando o valor limite está justamente no ponto central entre as duas observações localizadas na borda interna de cada classe de dados, as distâncias são iguais e a margem é a maior possível. Caso movamos o valor limite para um lado ou para outro, como a margem é a menor distância entre as observações e o valor limite, ela sempre será menor.

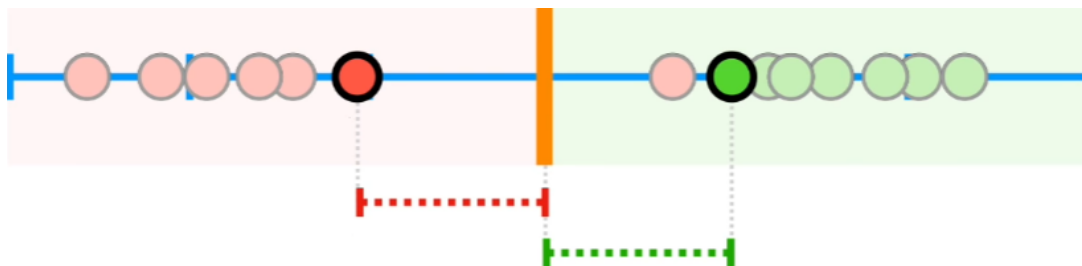
Vamos complicar um pouquinho mais os nossos dados de treinamento. Imagine agora que tenhamos um valor anômalo (*outlier*) dentro do meu conjunto de dados:



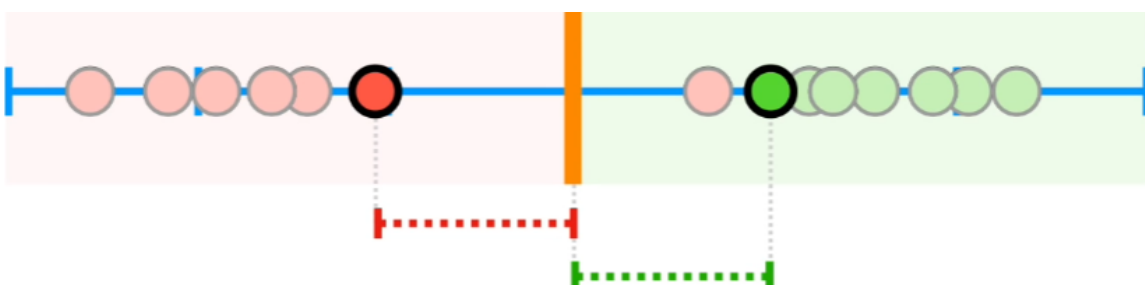
Note que o dado anômalo foi classificado como não obeso, mas está bem mais próximo da classe de obesos. O valor limite ficaria no ponto central entre as duas observações localizadas na borda interna de cada classe de dados conforme apresenta a imagem acima. Se tivermos uma nova observação como a representada pela bolinha preta na imagem seguinte, ela seria classificada como não obesa, muito embora esteja mais próxima da classe de obesos.

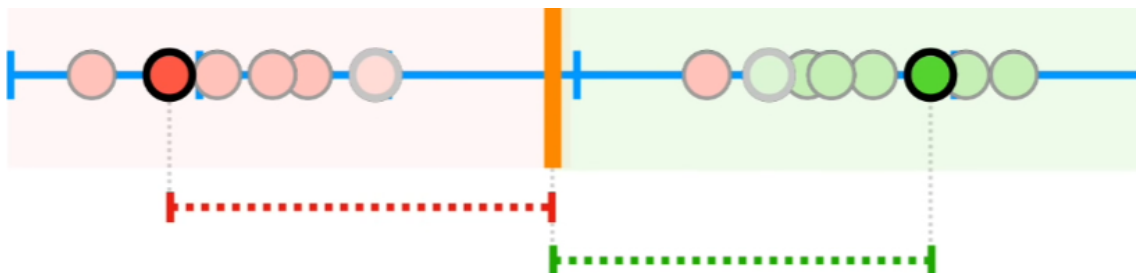


Podemos concluir que é complexa a atividade de escolher um limite quando o conjunto de dados de treinamento possui dados anômalos. A solução para esse tipo de problema é permitir classificações incorretas. No exemplo abaixo: se ignorarmos a bolinha vermelha da direita e utilizarmos um valor limite central entre as observações localizadas na borda interna de cada classe, nós vamos errar a classificação dos dados anômalos, mas classificaremos corretamente o restante.



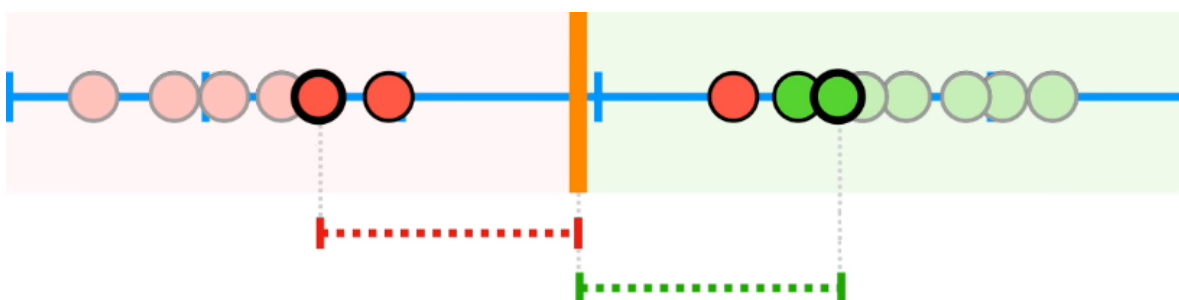
Quando nós permitimos classificações incorretas, a distância entre as observações e o valor limite é chamada de *margem flexível* (*soft margin*). Agora vejam as duas imagens a seguir:





Pergunto: *qual possui a margem flexível melhor?* Para responder a essa pergunta, utilizamos uma técnica chamada *cross-validation* (veremos adiante). Ela permite determinar quantas classificações incorretas e quantas observações devem ser permitidas dentro da margem flexível para conseguir a melhor classificação. Quando usamos uma margem flexível para determinar a localização ótima de um valor limite, estamos usando um Support Vector Classifier (SVC)¹⁴_[17].

Em tradução livre, seria chamado de Classificador de Vetores de Suporte. *E você sabe o que são os vetores de suporte?* São as observações dentro da margem flexível!



Todos esses exemplos são unidimensionais, mas poderíamos ter duas dimensões (Ex: Peso e Altura), três dimensões (Ex: Peso, Altura e Idade) ou mais dimensões. Quando temos apenas uma dimensão, o classificador de vetores de suporte é um ponto; quando temos duas dimensões, o classificador de vetores de suporte é uma linha; quando temos três dimensões, o classificador de vetores de suporte é um plano¹⁵_[18]; e quando temos mais dimensões, o classificador é um hiperplano.

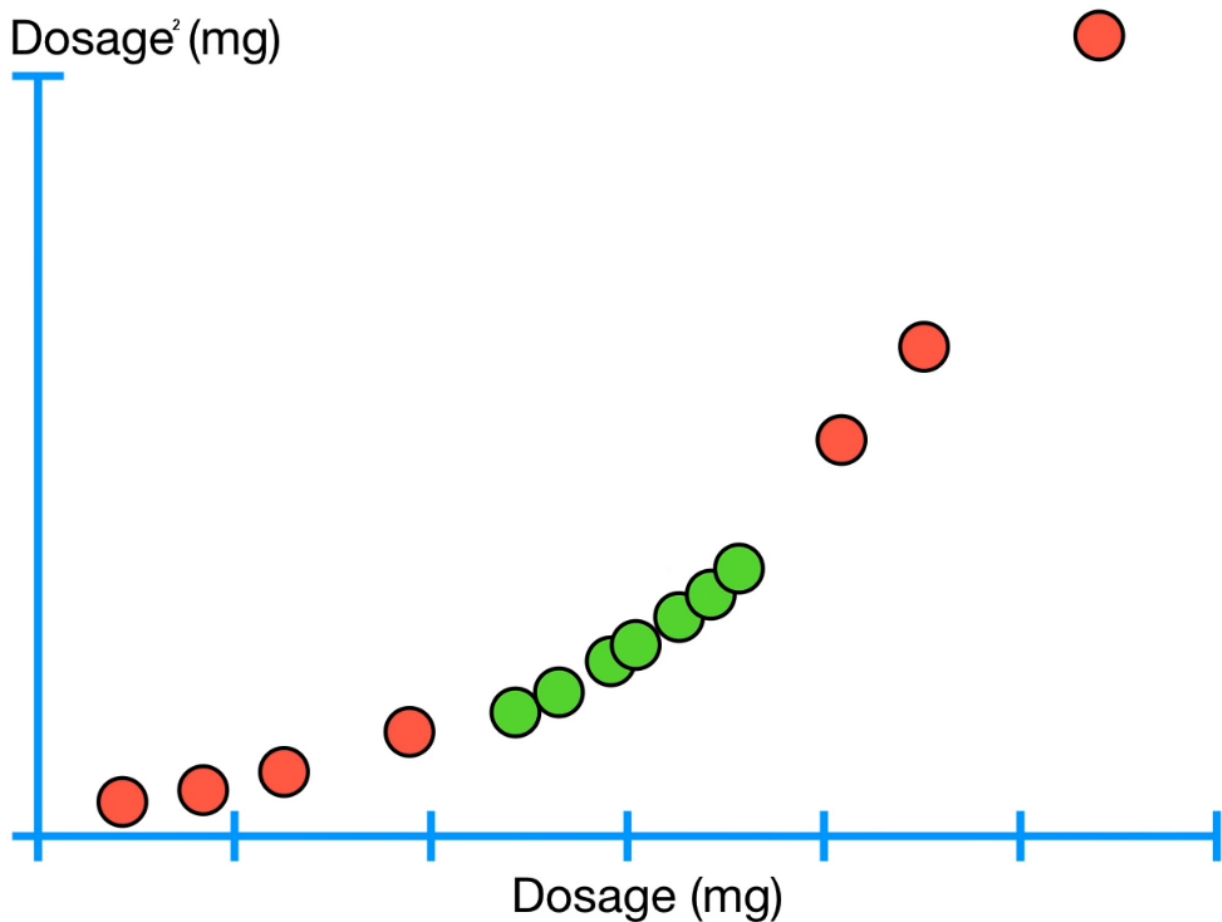
Agora vejam o conjunto de dados apresentado a seguir que representa a dosagem ideal de um remédio em miligramas. Há um ditado popular que diz: “*A diferença entre o remédio e o veneno é a dose*”. Isso faz sentido porque uma dose muito baixa de um remédio não faz o efeito esperado e uma dose muito alta pode até causar danos. Os pontos apresentados mostram mais ou menos isso: em vermelho, pacientes que não foram curados e em verde pacientes que foram curados.

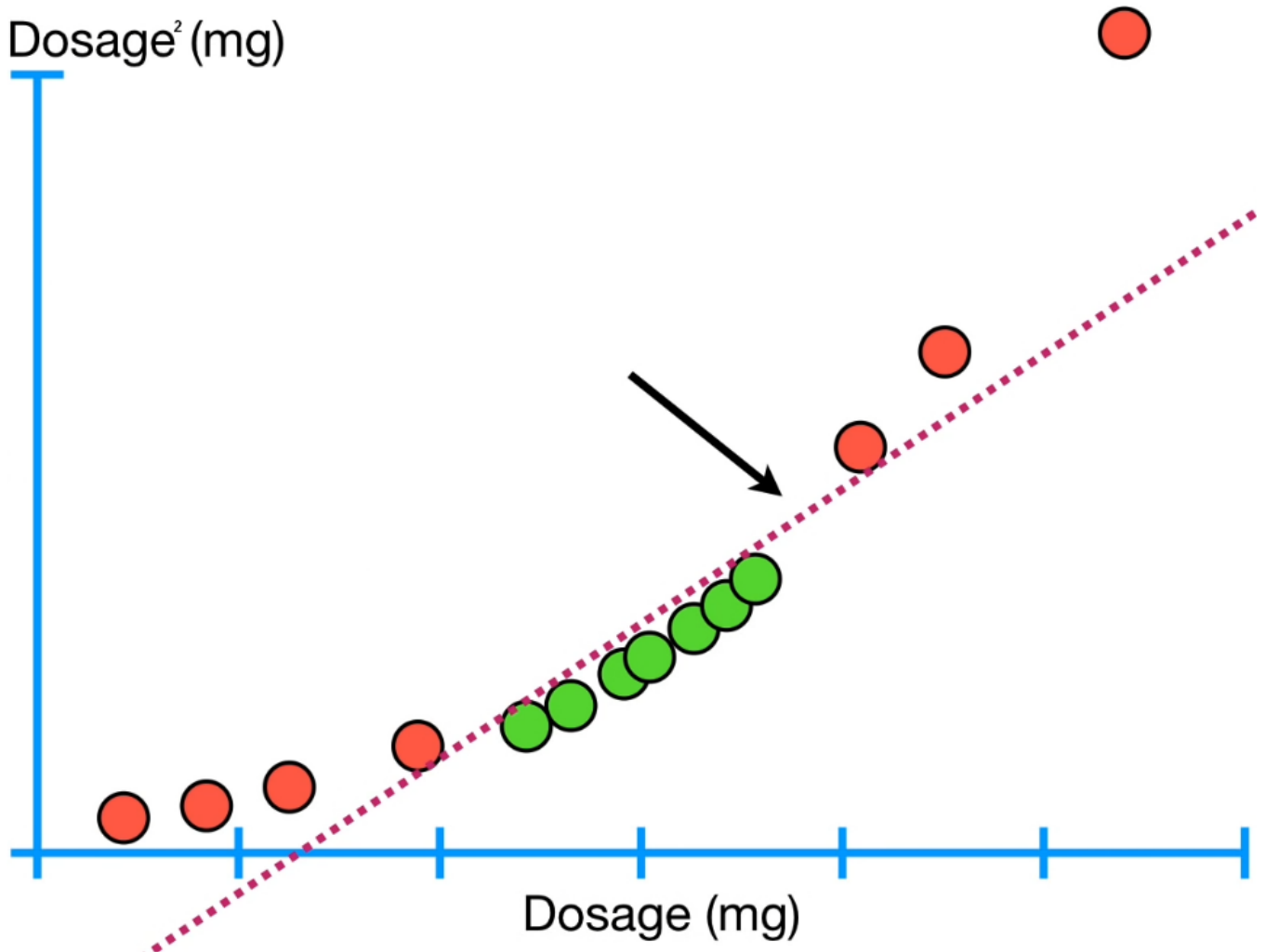


Podemos concluir que a dose não funciona caso seja pequena demais ou grande demais. E agora temos um problema: qualquer que seja a posição do valor limite, nós sempre teremos muitas classificações incorretas. Vocês podem tentar à vontade escolher um ponto na imagem acima que faça uma boa classificação dos dados que vocês não vão conseguir. Pode ser no início, no meio ou no fim, sempre haverá uma grande quantidade de classificações incorretas.

Pois é, os classificadores de vetores de suporte não possuem um bom desempenho para esse tipo de conjunto de dados. *E o que fazer agora, Diego?* Chegou o momento de ele brilhar: *Support Vector Machines* (SVM). Ele resolve o problema de selecionar um valor limite que permita classificar com acurácia um conjunto de dados por meio da adição de dimensões ao problema, isto é, o que era um problema unidimensional agora será um problema bidimensional.

Vejamos um exemplo: vamos construir um plano (bidimensional) em que o eixo X continuará sendo a dosagem do remédio e o eixo Y será a dosagem ao quadrado:





Viram a mágica? Agora se um novo dado surgir, será possível classificá-lo de forma mais satisfatória. Em síntese, a ideia por trás desse algoritmo é: iniciar com dados em uma dimensão relativamente baixa (Ex: uma dimensão), depois mover os dados para uma dimensão maior (Ex: duas dimensões), e por fim encontrar um classificador de vetores de suporte que consiga separar os dados de dimensão maior em dois grupos.

E por que nós usamos dosagem ao quadrado? Não poderia ser outra dimensão? O SVM utiliza um conceito chamado Funções de Kernel (ou simplesmente Kernel). Essas funções permitem transformar dados de entrada no formato necessário ao encontrar classificadores de vetores de suporte em dimensões maiores, tais como funções lineares, polinomiais ou radiais. *E quais são as vantagens e desvantagens desse algoritmo?*

Como vantagens, podemos afirmar que ele é bastante efetivo quando as classes dos pontos de dados estão bem separadas; é efetivo até em espaços de alta dimensionalidade; é bastante eficiente nos casos em que o número de dimensões (variáveis) é maior que o número de amostras (linhas de uma tabela); é relativamente eficiente em relação ao usuário de memória computacional; e trabalha bem com imagens.

Como desvantagens, podemos afirmar que ele não é ideal para grandes conjuntos de dados; não tem um bom desempenho quando o conjunto de dados possui muito ruído ou quando as classes de dados não são bem separadas; é desafiador escolher uma função de kernel ótima; em grandes bases de dados, ele precisa de mais treinamento; como não se trata de um modelo

probabilístico, não é possível explicar a classificação por esses termos; e é mais difícil de interpretar que outros métodos.

Por fim, é importante saber que o SVM é considerado um método não paramétrico¹⁶ [19], dado que não existe uma função de mapeamento entre dados de entrada e dados de saída definida por um conjunto finito de números (chamados de parâmetros) que plenamente caracterizam todos os possíveis resultados da função de mapeamento e cujos valores específicos são determinados durante o treinamento. *E as funções de kernel?*

As funções de kernel são escolhidas pelo usuário e não otimizadas pelo processo de treinamento, logo elas são consideradas hiperparâmetros (não é necessário entender isso agora). Além disso, podemos afirmar que o SVM é sensível à escala dimensional dos conjuntos de dados. *Como assim, Diego?* Isso significa que mudanças na escala do conjunto de dados afetam as previsões de um modelo de aprendizado de máquina.

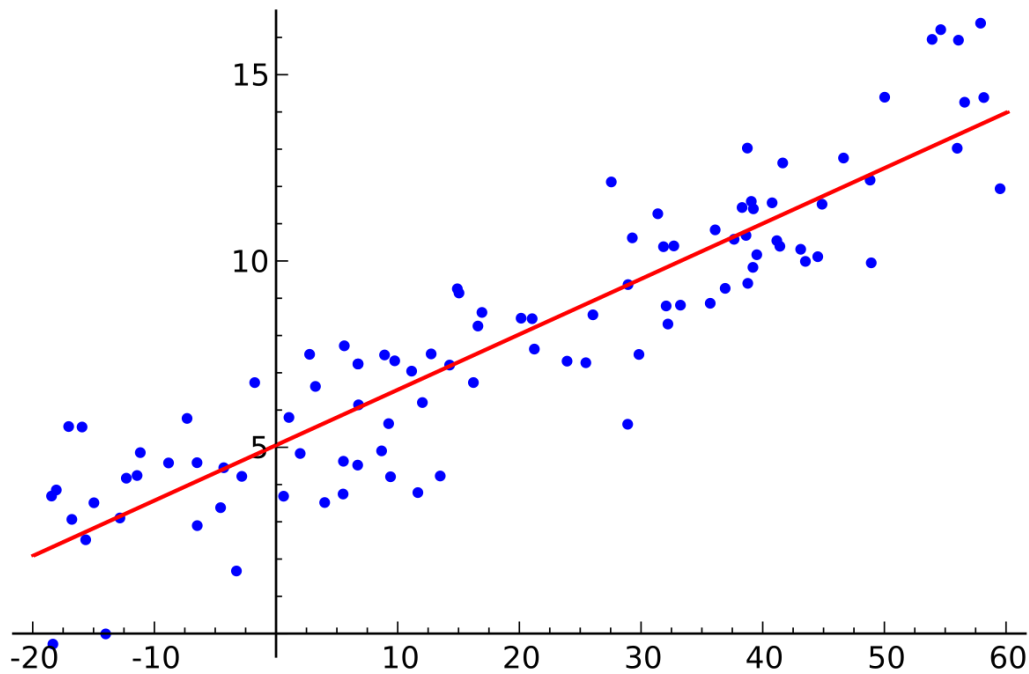
Veja que estamos falando de mudanças na escala do conjunto de dados e, não, de mudanças nas dimensões do conjunto de dados. Como dados podem vir em diferentes tamanho e formatos, é importante realizar o pré-processamento (transformação, padronização ou normalização) para que eles fiquem com dimensões razoáveis de forma que o algoritmo possa definir bem a fronteira dos dados por meio dos vetores de suporte (Ex: escalar cada atributo entre [0,1] ou [-1,1]).

Regressão

INCIDÊNCIA EM PROVA: baixa

Galera, um parente muito próximo da classificação é a regressão. **Na regressão, em vez de prever uma categoria, o objetivo é prever um número.** Vamos pegar o exemplo da Target novamente! Eles queriam saber não apenas se cada cliente estava grávida, mas quando enviar cada cupom de desconto. Então eles conseguiram estimar as datas de nascimento também dos bebês. Essa é uma questão de regressão, i.e., quantas semanas até a cliente dar à luz.

A Regressão depende muitas vezes de dezenas ou mesmo milhares de variáveis ou características que descrevam cada exemplo e encontra uma equação ou curva para ajustar os pontos de dados. Como na classificação, muitas técnicas de regressão dão a cada característica um peso, então combinam contribuições positivas e negativas dos recursos ponderados para obter uma estimativa. Vejam no exemplo a seguir:



Notem que baseado em diversos pontos de dados, foi traçada uma linha capaz de estimar dados – poderia ter sido não-linear também. Galera, essa linha do gráfico é dada por uma equação baseada em variáveis. Logo, se eu tenho a equação, eu posso estimar qualquer outro valor – basta jogar na equação e esperar o resultado. Vamos supor que o exemplo acima seja dado pela equação $y = 2,2923x - 46,244$ – se eu quiser saber qualquer valor de y , é só mudar o valor de x .

E, assim como a classificação, a regressão também é usada em vários lugares. Um dos exemplos mais conhecidos é o Google Trend da Gripe. Em 2008, o Google começou a publicar estimativas em tempo real de quantas pessoas teriam gripe com base em pesquisas por palavras como "febre" e "tosse". **Em alguns casos, ele foi capaz de prever surtos regionais de gripe até 10 dias antes de serem notificados pelo CDC (Centros de Controle e Prevenção de Doenças).**

Em 2010, o CDC identificou um pico de casos de gripe na região do Atlântico dos Estados Unidos. No entanto, os dados das consultas de pesquisa do Google sobre os sintomas da gripe conseguiram mostrar esse mesmo pico duas semanas antes do relatório do CDC! Inicialmente, o Google tinha uma precisão de 97% em relação ao CDC, porém em anos subsequentes ele reduziu sua precisão e o Google decidiu retirar do ar enquanto não houvesse uma precisão melhor.

<https://www.google.org/flutrends> [20]

Em uma definição formal, Navathe afirma que a regressão é uma aplicação especial da regra de classificação. Se uma regra de classificação é considerada uma função sobre variáveis que as mapeia em uma classe destino, a regra é chamada regressão. **Isto ocorre quando, ao invés de mapear um registro de dados para uma classe específica, o valor da variável é previsto (calculado) baseado em outros atributos do próprio registro.** Bacana?

Regressão Linear

INCIDÊNCIA EM PROVA: baixíssima

A regressão linear é um tipo de algoritmo de aprendizado de máquina supervisionado utilizado na mineração de dados. Ela é usada para prever uma variável de destino contínua ajustando uma equação linear aos pontos de dados. Baseia-se na relação entre as variáveis independentes (preditoras) e a variável dependente (alvo). O algoritmo de regressão linear encontra a melhor linha de ajuste que minimiza a soma dos erros quadrados.

Essa linha de melhor ajuste é então usada para prever valores para a variável dependente. Dito de outra forma, a ideia é modelar o relacionamento entre uma variável dependente (Y) e uma ou mais variáveis independentes (X). Esse modelo é utilizado para fazer previsões sobre a variável dependente ajustando uma equação linear aos dados observados. A equação de regressão linear mostra a relação linear entre a(s) variável(is) independente(s) e a variável dependente.

Sendo assim, o modelo de regressão para uma única variável preditora x, ou regressão linear simples, pode ser definido pela seguinte equação da reta:

$$y = a + bx$$

em que a e b são coeficientes de regressão (pesos) e especificam o intercepto do eixo y e a inclinação da reta, respectivamente.

Deve-se, então, encontrar valores para os coeficientes de regressão, de forma que a reta (ou o plano/hiperplano, se for uma regressão linear multivariada) se ajuste aos valores assumidos pelas variáveis nos exemplares de um conjunto de dados. O melhor ajuste da reta pode ser encontrado, por exemplo, pelo método dos mínimos quadrados, o qual minimiza o erro entre os valores das variáveis nos exemplares do conjunto de dados e os valores estimados pelo regressor.

Regressão Logística

INCIDÊNCIA EM PROVA: baixíssima

Na mineração de dados, a regressão logística é uma técnica de modelagem preditiva utilizada para problemas de classificação. Trata-se de um algoritmo de aprendizado supervisionado que usa um conjunto de dados de treinamento rotulados para construir um modelo que pode prever com precisão os resultados de dados não vistos. A regressão logística é usada para identificar padrões em grandes conjuntos de dados e para estimar a probabilidade de um determinado evento ocorrer.

Análise de Agrupamentos

INCIDÊNCIA EM PROVA: média

Análise de agrupamentos (também chamados de clusters, grupos, aglomerados, segmentos, partições ou agregações) é uma técnica que visa fazer agrupamentos automáticos de dados segundo o seu grau de semelhança, permitindo a descoberta por faixa de valores e pelo exame de atributos das entidades envolvidas. **Como o nome sugere, o objetivo é descobrir diferentes clusters em uma massa de dados e agrupá-los de uma forma que ajude com sua análise.**

Um agrupamento é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação uma vez não necessita que os registros sejam previamente categorizados – trata-se de um aprendizado não-supervisionado. Além disso, ela não tem a pretensão de classificar, estimar ou prever o valor de uma variável, ela apenas identifica os grupos de dados similares.

Vamos imaginar um site: www.mercadolivre.com.br [21]! Lá existem milhões de produtos de absolutamente tudo que você imaginar – mesmo dentro de uma única categoria, existem uma infinidade de produtos diferentes. **Por essa razão, o Mercado Livre organiza as coisas em subcategorias, mas é inviável ter um funcionário que fique organizando todos os anúncios em categorias, subcategorias, entre outros.**

Em vez disso, a empresa pode usar técnicas de agrupamento para agrupar automaticamente os produtos. Mais uma vez: cada produto deve primeiro ser dividido em características numéricas, como quantas vezes a palavra “impressora” aparece na descrição ou quem é a fabricante. **O método de cluster mais simples é adivinhar quantas subcategorias distintas devem existir.** Sim, você dá um “chute” de categorias...

Em seguida, você agrupa itens aleatoriamente em várias categorias diferentes e depois continua mudando itens entre categorias para tornar cada o grupo mais preciso. No final, acreditem ou não, produtos similares acabam se agrupando, mas não precisamos parar por aí! Imaginem dois anúncios de um mesmo modelo de câmera, mas com cores diferentes! Eles não precisam ficar em categorias separadas, porque são apenas variantes do mesmo produto.

Dessa forma, além das subcategorias, seria interessante mesclar alguns grupos. Sites como o Mercado Livre fazem isso por meio de uma técnica chamada clustering hierárquico. **Em vez de um único conjunto de categorias, o agrupamento hierárquico produz uma espécie de árvore taxonômica, aglomerando ou dividindo elementos.** *Adivinhem quem utilizou essa*

técnica? Cambridge Analytica. Ela agrupou eleitores que respondiam ao mesmo tipo de publicidade!

Em síntese, essa técnica realiza agrupamentos comuns chamados *clusters*, de modo que o grau de associação seja forte entre os membros do mesmo grupo e fraco entre membros de grupos diferentes. Bacana? Antes de falarmos sobre os principais algoritmos de agrupamento nos tópicos seguintes, vamos primeiro ver algumas classificações. Iniciemos pela classificação em determinístico e estocástico:

CLASSIFICAÇÃO	DESCRIÇÃO
Determinístico	Métodos determinísticos apresentam sempre o mesmo agrupamento, independente de parâmetros do algoritmo e/ou da condição inicial.
estocástico	Métodos estocásticos podem apresentar diferentes soluções dependendo dos parâmetros e/ou da condição inicial.

Os algoritmos de agrupamento também podem ser divididos em algoritmos hierárquicos ou algoritmos particionais:

CLASSIFICAÇÃO	DESCRIÇÃO
HIERÁRQUICO	Métodos hierárquicos criam uma decomposição hierárquica dos dados, podendo ser divididos em aglomerativos ou divisivos, baseados em como o processo de composição/decomposição é efetuado. Métodos aglomerativos começam com cada objeto pertencendo a um grupo e unem sucessivamente objetos. Métodos divisivos começam com todos os objetos fazendo parte do mesmo grupo e particionam sucessivamente os grupos em grupos menores.
PARTICIONAL	Dado um conjunto com n objetos, um método particional constrói k partições dos dados, sendo que cada partição representa um cluster ($k \leq n$). Dado o número k de partições, um método particional cria uma partição inicial e emprega um algoritmo de realocação iterativa que tem por objetivo melhorar o particionamento movendo objetos entre grupos.

Por fim, os algoritmos particionais podem ser divididos em algoritmos rígidos (*hard*), algoritmos suaves (*soft*) ou algoritmos difusos (*fuzzy*).

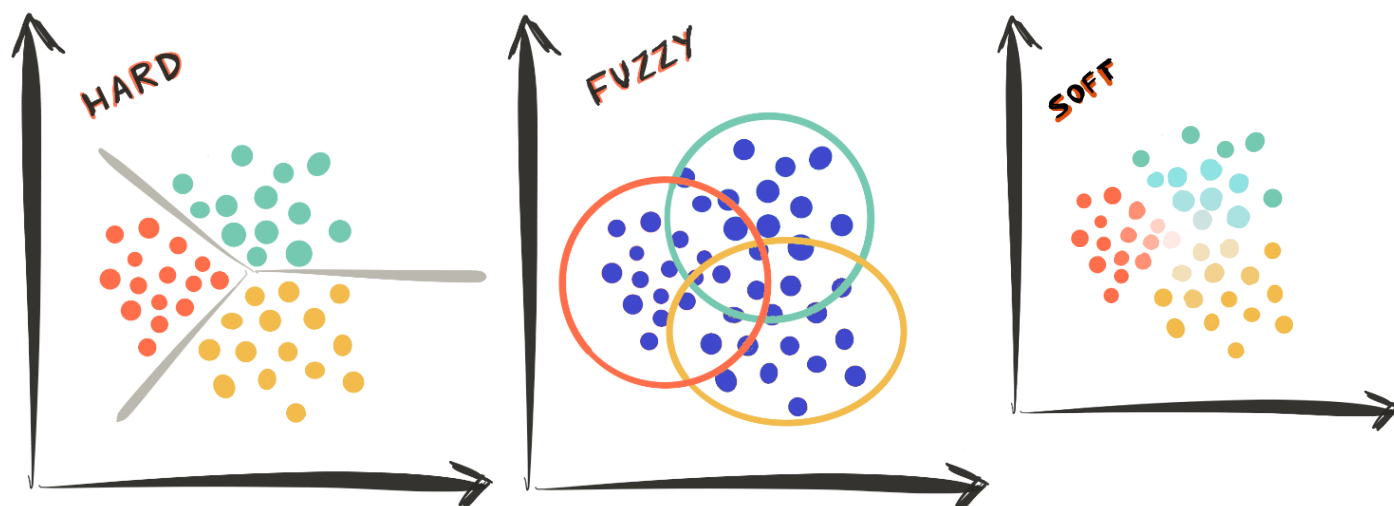
CLASSIFICAÇÃO	DESCRIÇÃO
Rígidos (<i>hard</i>)	Em agrupamentos rígidos, cada objeto pertence a um único grupo.
Suaves (<i>soft</i>)	Em agrupamentos suaves, cada objeto pertence completamente a mais de um grupo.
Difusos (<i>fuzzy</i>)	Em agrupamentos difusos, cada objeto pertence parcialmente a mais de um grupo.



Vamos supor que em nosso *dataset*, queremos agrupar um conjunto de mangas. Ao utilizar um método de agrupamento rígido, um possível agrupamento poderia ser: manga verde, manga amarela e manga vermelha. Ao utilizar um método de agrupamento suave, poderíamos ter esses mesmos rótulos, mas baseado no grau de pertencimento (calculado por meio de probabilidade) aos grupos (Ex: Manga 1: 45% Verde, 40% Vermelha, 15% Amarela, logo será considerada Verde).

E ao utilizar um método de agrupamento difuso, poderíamos ter esses mesmos rótulos, mas uma manga pode ser considerada simultaneamente (Verde e Vermelha) ou (Verde e Amarela) ou (Amarela e Vermelha) ou (Verde, Vermelha e Amarela) de acordo com seus graus de pertencimento. *Então qual é a diferença para o anterior, professor?* No suave, apesar dos graus de pertencimento, cada objeto tem um único rótulo; no difuso, o objeto pode ter mais de um rótulo.

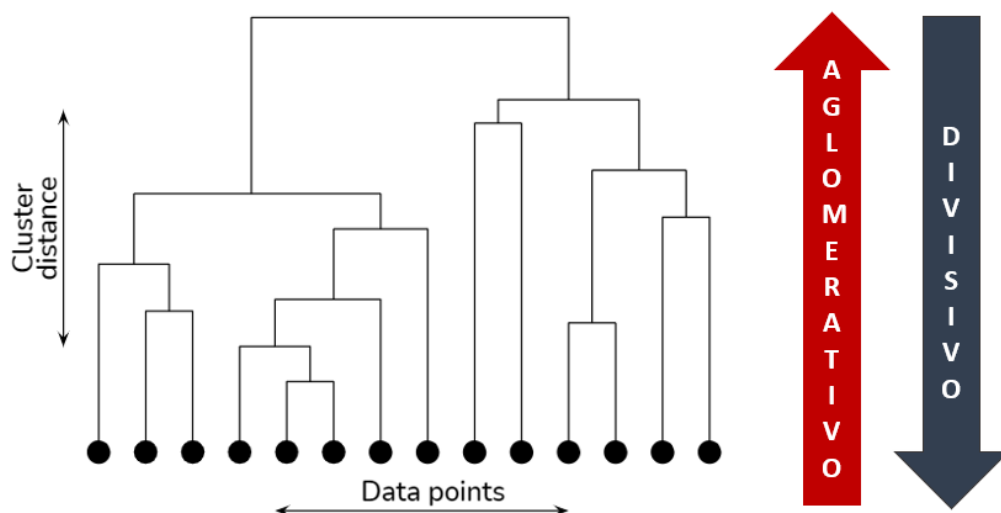
Nas três imagens seguintes, eu tentei representar esses três cenários em um plano cartesiano (eu fiz esse desenho no iPad, então não ficou muito bom, mas é possível de entender).



Agrupamento Hierárquico

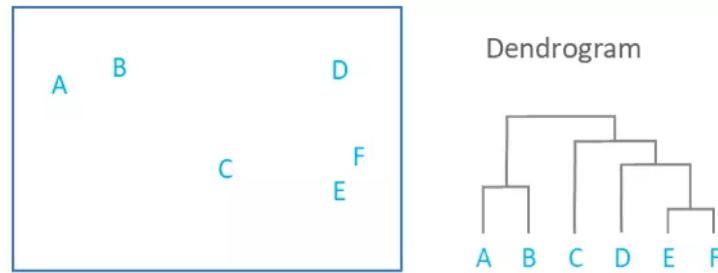
INCIDÊNCIA EM PROVA: baixíssima

O conceito de agrupamento hierárquico se baseia na construção e análise de um dendrograma. *O que seria isso, professor?* **É basicamente um diagrama que exhibe a relação hierárquica entre objetos.** No dendrograma apresentado a seguir, a parte de cima representa as gerações mais antigas (avós) e a parte de baixo representa as gerações mais novas (filhos), contendo no meio gerações intermediárias.

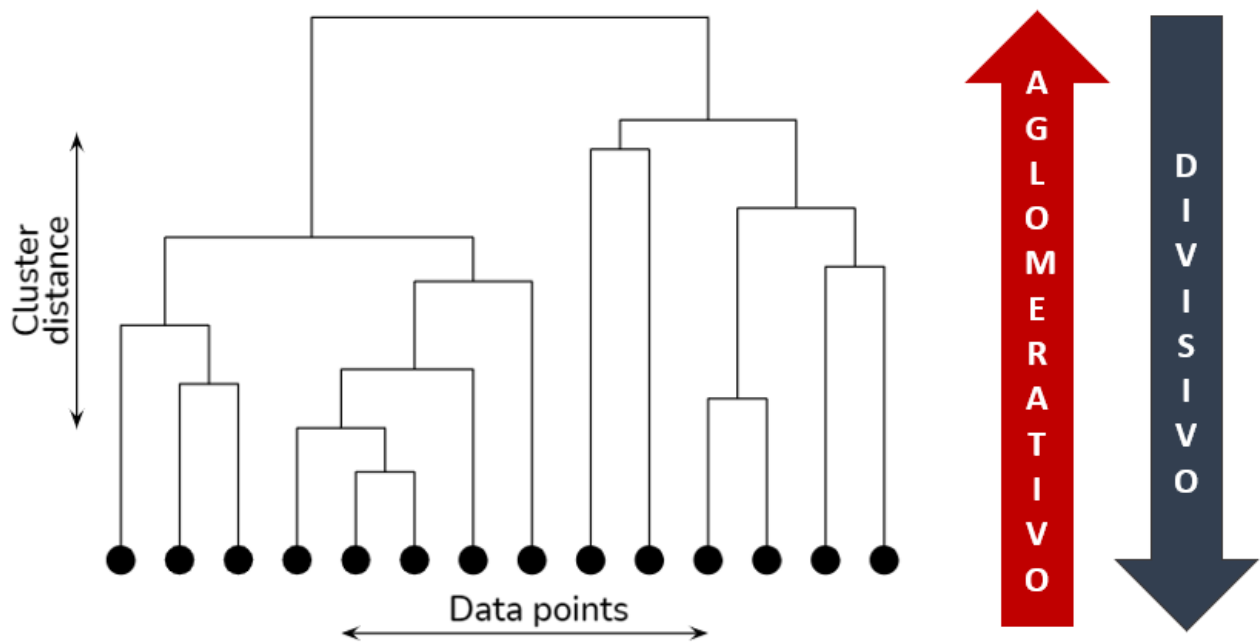


Para interpretar um dendrograma, é preciso entender que no eixo das abscissas (x) temos os pontos de dados e no eixo das ordenadas (y) temos a distância entre os grupos. No exemplo a seguir, isso fica mais claro: os pontos mais próximos são E e F, logo notem que eles possuem uma altura menor; em seguida, os pontos mais próximos são A e B, logo notem que ele tem a

altura um pouco maior; depois o ponto mais próximo é entre o grupo (E,F) e D; depois (E,F,D) e C; depois (E,F,D,C) e (A,B).



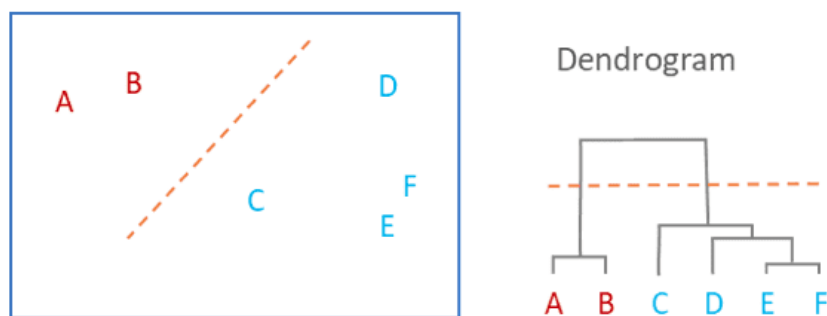
Vocês notaram que nós construímos o dendrograma de baixo para cima (*bottom-up*)? Pois é, utilizamos um método aglomerativo! Se fosse de cima para baixo (*top-down*), seria divisivo.



O Método Aglomerativo é também chamado de Método AGNES (*Agglomerative Nesting*). Conforme vimos, inicialmente cada objeto é considerado um grupo de um único elemento (folha) e, a cada passo do algoritmo, dois grupos similares são combinados para formar um grupo maior (nós). Esse procedimento ocorre iterativamente de baixo para cima até que todos os pontos sejam membros de um único grupo (raiz).

O Método Divisivo é também chamado de Método DIANA (*Divisive Analysis*). Conforme vimos, trata-se do inverso do método anterior: ele começa inicialmente com a raiz, na qual todos os objetos são incluídos em um único grupo. A cada passo da iteração, os grupos mais heterogêneos são divididos em dois. O processo permanece iterando até que todos os objetos sejam seu próprio cluster de um único objeto.

Por fim, após terminar a construção do dendrograma, podemos escolher um ponto da árvore para realizar a partição dos grupos conforme é exibido na imagem seguinte. De acordo com Matt Harrison [Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python]: “Após terminar a construção, tem-se um dendrograma, isto é, uma árvore que controla quando os clusters foram criados e qual é a métrica das distâncias”.



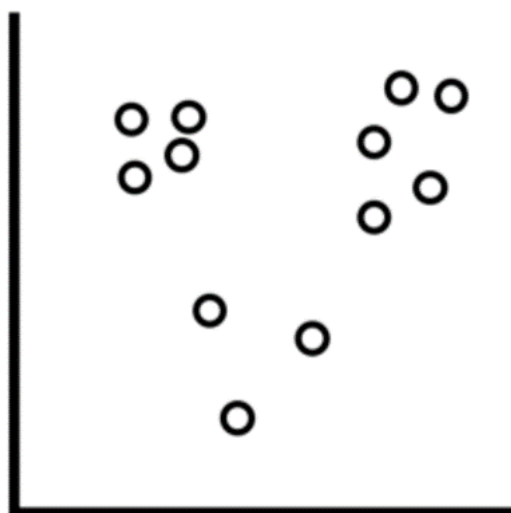
Essa definição já foi cobrada em prova em sua literalidade, mas está errada! Dendrograma é apenas uma representação visual do resultado do processo de agrupamento e não controla coisa alguma.

K-Médias

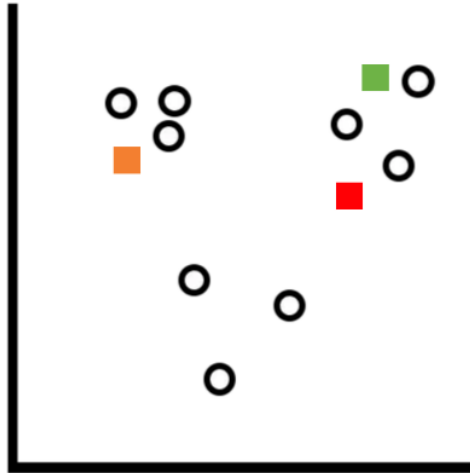
INCIDÊNCIA EM PROVA: baixíssima

Também chamado de K-Means, trata-se de algoritmo de agrupamento que basicamente agrupa dados em k grupos, em que k é um valor arbitrário definido pelo usuário. Esse algoritmo busca minimizar a soma de todos os quadrados das distâncias entre os pontos de dados e um ponto chamado centroide (também conhecido como semente). Galera, é um pouco complicado explicar de forma satisfatória esse algoritmo em texto, mas eu fiz uns desenhos no Paint para ajudar

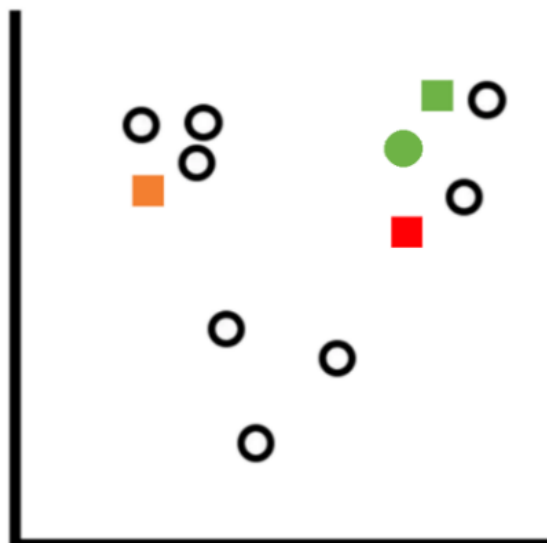
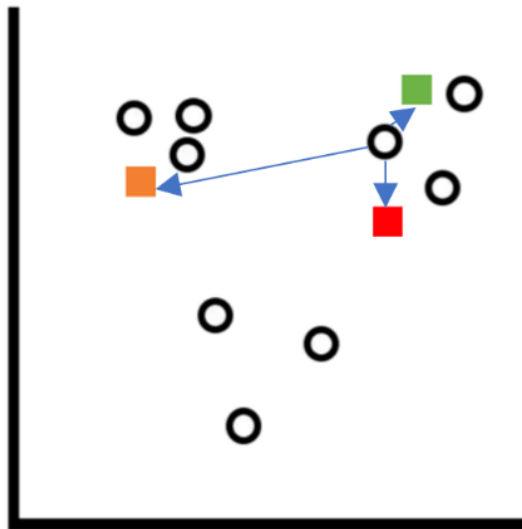
O primeiro passo desse algoritmo é reunir os dados e determinar em quantos grupos nós queremos dividi-los. Em nosso exemplo, eu vou querer dividir em três grupos ($k = 3$).



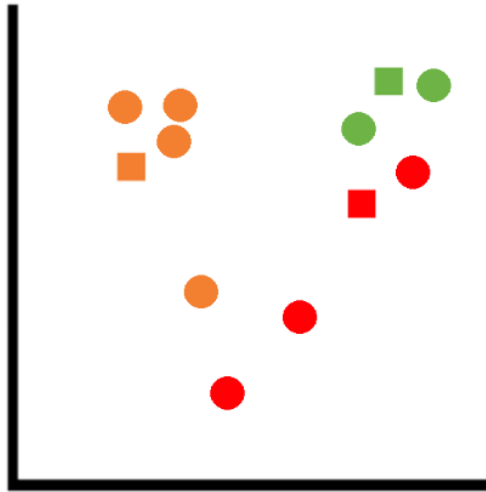
No segundo passo, nós selecionamos k pontos aleatórios. Como nós decidimos no passo anterior que $k = 3$, então vamos selecionar três pontos no *dataset* (esses pontos são os centroides).



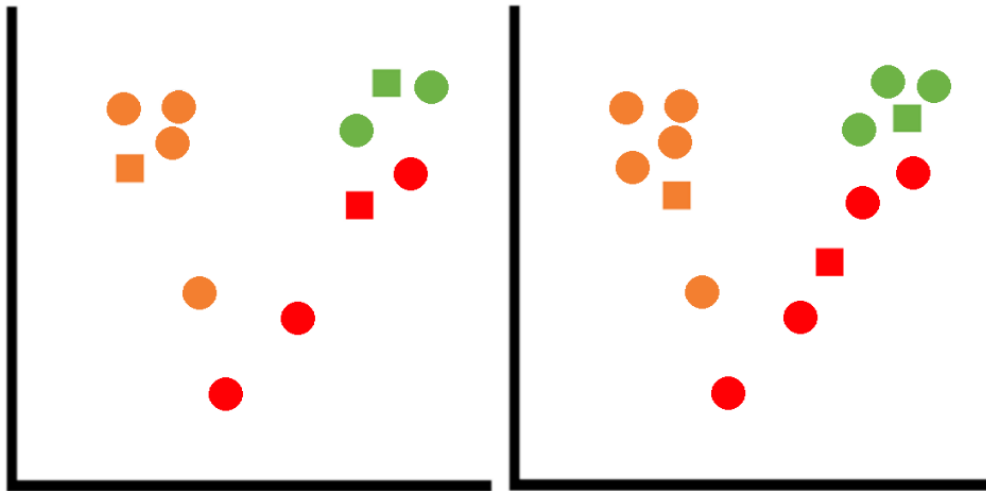
No terceiro passo, nós vamos calcular a distância de cada um dos pontos para cada um dos centroides. Por exemplo: começando pelo ponto lá no canto superior direito conforme mostra a imagem seguinte. Note que eu inseri três setas azuis saindo desse ponto em direção aos três centroides. Ao calcular a distância, é possível ver que o ponto verde é o mais próximo, logo esse primeiro ponto analisado será marcado como verde.



Agora faltam 8 pontos! Seguimos para o próximo e repetimos esse procedimento para os pontos restantes até chegar ao resultado apresentado a seguir:



No quarto passo, nós temos que identificar novos centroides. Para tal, nós vamos migrá-los de suas posições iniciais para a posição mais próxima dos centros dos pontos da sua respectiva cor. *Diego, segura a onda aí que foi rápido demais! Vamos lá: para cada ponto de um determinado grupo (laranja, vermelho e verde), nós vamos somar o valor de suas coordenadas e dividir pela quantidade de pontos (por isso é uma média). Essa coordenada resultante será a nova posição do centroide*



O quinto passo trata basicamente de repetir o processo até o momento em que os centroides não alteram mais suas posições (ou a alteração é minúscula) e as observações não mudam mais de grupo, alcançando a posição ótima¹⁷ [22]. Lembrando que esse algoritmo gera agrupamentos em função da distância euclidiana (soma do quadrado das diferenças das coordenadas) de um ponto central. É importante dizer que o algoritmo k-means não é determinístico. *Como assim, Diego?*

Isso significa que a escolha aleatória inicial dos centroides influencia no agrupamento gerado, logo – a cada execução do algoritmo com diferentes escolhas iniciais de centroide – podemos obter resultados distintos. O k-Means é um algoritmo rápido, eficiente, fácil de entender e implementar, no entanto a escolha do valor ótimo de k é desafiadora, a escolha do centroide inicial influencia no resultado final e ele é bastante sensível a anomalias e ruídos.

K-Medoides

INCIDÊNCIA EM PROVA: baixíssim

Em primeiro lugar: *o que é um medoide?* Medoide é o objeto com menor dissimilaridade média em relação a todos os outros objetos, isto é, aquele mais centralmente localizado de um grupo. Nós acabamos de ver que os centroides não são pontos reais, mas simplesmente a média dos pontos presentes em um agrupamento. A ideia do k-Medoides é fazer com que os centroides finais sejam pontos de dados reais.

Em outras palavras, o k-Medoides escolhe objetos da própria base como protótipo¹⁸ [23] em lugar de um centroide, ao passo que o k-Médias calcula o centro do grupo a partir dos objetos contidos neles. Trata-se de um algoritmo mais resistente a ruídos e valores anômalos que o k-Médias, dado que o centro de um agrupamento será um objeto da própria base. Dessa forma, ruídos e valores anômalos influencia menos a definição do centro.

Fuzzy K-Médias

INCIDÊNCIA EM PROVA: baixíssima

O Fuzzy K-Médias é basicamente uma extensão do k-Médias em que cada objeto pode pertencer a mais de um grupo. Ele se baseia na ideia de que cada objeto possui um grau de pertencimento em relação a cada um dos agrupamentos. O grau de pertencimento é um valor normalizado entre 0 e 1, mas não representam probabilidades, logo a soma dos valores pode ser maior que 1. De resto, esse algoritmo é bastante similar ao K-Médias.

Ele tem alguns problemas: está sujeito a uma localização ótima do centroide, o resultado também depende da condição inicial e o valor de k deve ser definido a priori.

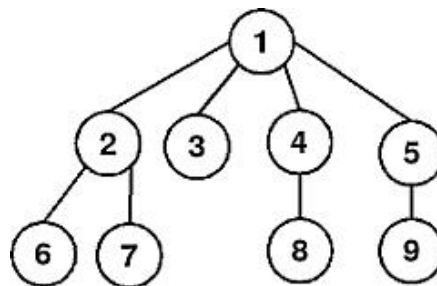
Árvore Geradora Mínima

INCIDÊNCIA EM PROVA: baixíssima

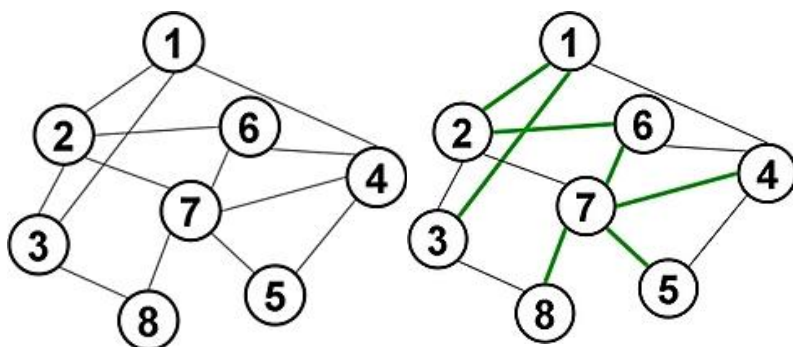
A Árvore Geradora Mínima (Minimal Spanning Tree – MST) é um algoritmo utilizado em diversas áreas de tecnologia da informação. Em nosso contexto, trata-se de um método baseado em teoria dos grafos utilizado para segmentar um conjunto de dados em diferentes grupos. *Qual é o diferencial dela?* Bem, nos algoritmos anteriores, partia-se sempre de um conjunto inicial de protótipos e havia um processo iterativo de alocação de objetos aos protótipos.

Lembrando que os protótipos são aqueles “chutes” iniciais (centroide no k-Médias e medoide no k-Medoide). Esse processo ocorria repetidamente, calculando novos protótipos a fim de particionar um conjunto de dados em k grupos e minimizar uma função de custo proporcional à distância intragrupo. Aqui é diferente: não há definição de protótipos para segmentar a base de dados em diferentes grupos.

Podemos afirmar que uma árvore é dita geradora se ela interliga (direta ou indiretamente) todos os nós de um grafo¹⁹ [24], ou seja, se todos os nós do grafo fazem parte da árvore - não ficou nenhum nó “isolado”. Perceba que, por exemplo, os nós 1 e 8 da árvore seguinte não estão diretamente ligados (não existe aresta entre eles), mas mesmo assim ambos fazem parte da árvore. Se não existisse a aresta que liga 4 e 8, o nó 8 ficaria “sozinho”, isolado do resto da estrutura.

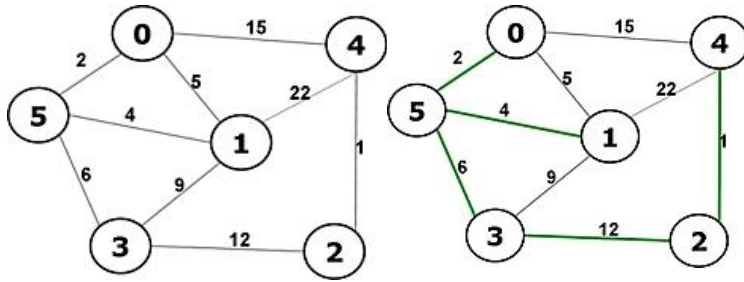


Os nós {1,2,3,4,5,6,7,9} continuariam a formar uma árvore, mas essa árvore não seria mais geradora, pois não inclui todos os nós do grafo - o nó 8 ficou de fora. A imagem da esquerda logo abaixo apresenta um grafo qualquer com 8 nós e 13 arestas. Vamos escolher somente algumas dessas arestas (o menor número possível), de forma que o grafo fique conectado. Na imagem da direita, as arestas escolhidas estão destacadas.



Não precisamos escolher nenhuma aresta a mais, já que todos os nós já estão conectados. Também não podemos escolher nenhuma aresta a menos, pois estaríamos desconectando os nós. Perceba como escolhemos justamente uma árvore geradora - um grafo sem ciclos (árvore)

que conecta todos os nós (geradora). Lembrando que esse é apenas um exemplo, mas existiam outras escolhas de arestas para chegar ao mesmo objetivo.



Dito isso, agora nós queremos chegar à árvore geradora mínima! Para tal, vamos levar em consideração o peso das arestas - queremos escolher, entre as árvores geradoras possíveis, aquela que possua o menor peso total, onde o peso total é dado pela soma dos pesos das arestas da árvore. Vejamos um exemplo: na esquerda, temos um grafo qualquer com 6 nós e 9 arestas (com seus respectivos pesos).

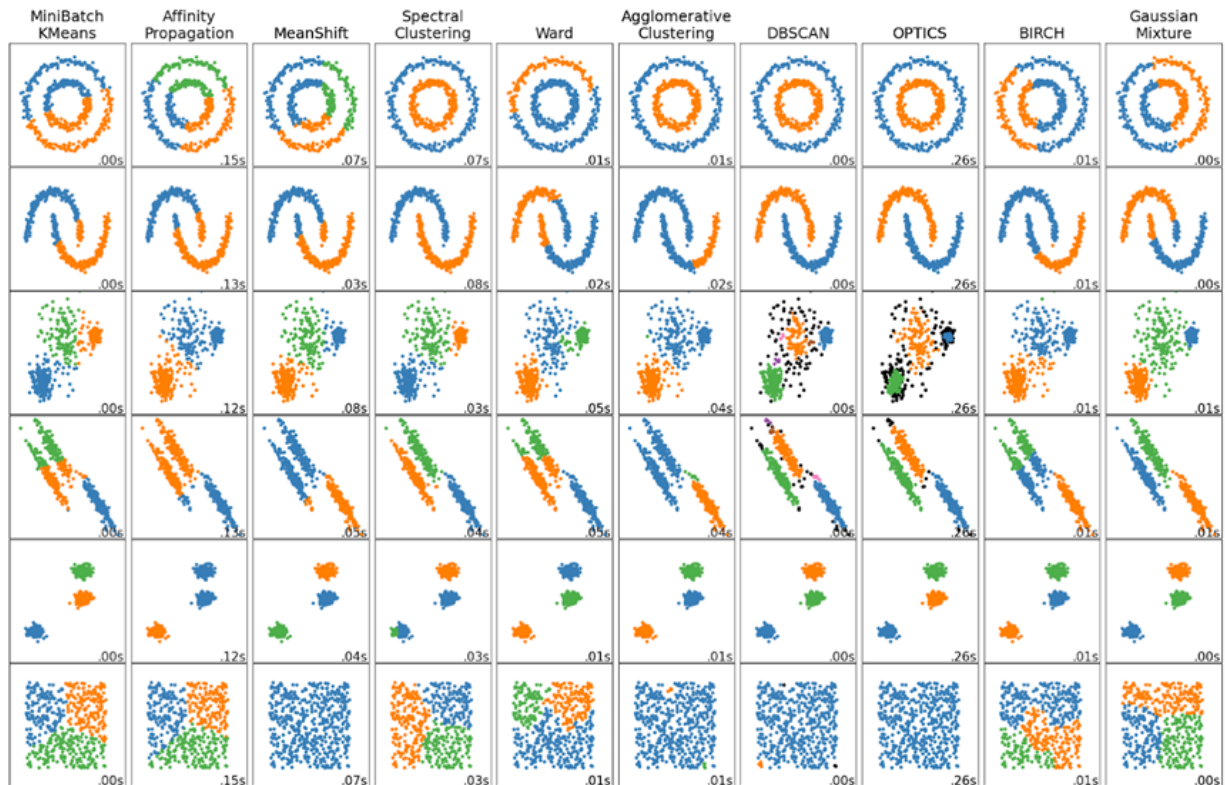
Na direita, temos as arestas escolhidas que formam uma árvore geradora de peso $1+2+4+5+12 = 25$. Como essa é a árvore geradora que gera o menor peso, ela é chamada de árvore geradora mínima. Existem diversos algoritmos que permitem calcular qual é a árvore geradora mínima, mas não vamos entrar nesses detalhes. É importante também saber que o peso da aresta pode representar qualquer valor em um problema real, como custo, fluxo, confiabilidade, tempo, entre outros.

Em nosso contexto, a Árvore Geradora Mínima permite identificar os limites de valores de um conjunto de dados que permitem dividir essa base de dados em grupos (*clusters*).

DBSCAN

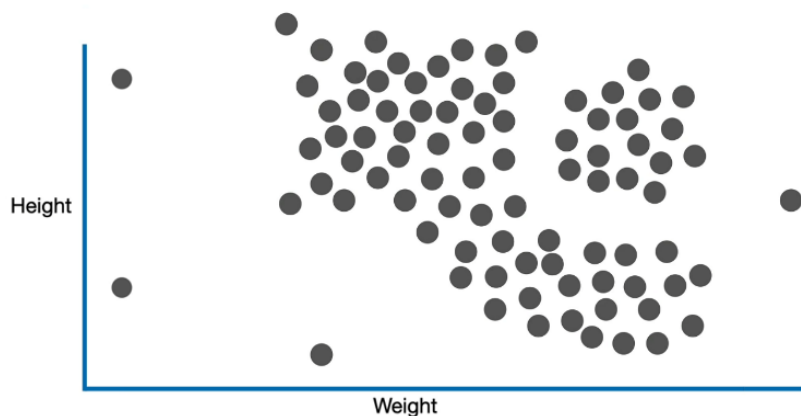
INCIDÊNCIA EM PROVA: baixíssima

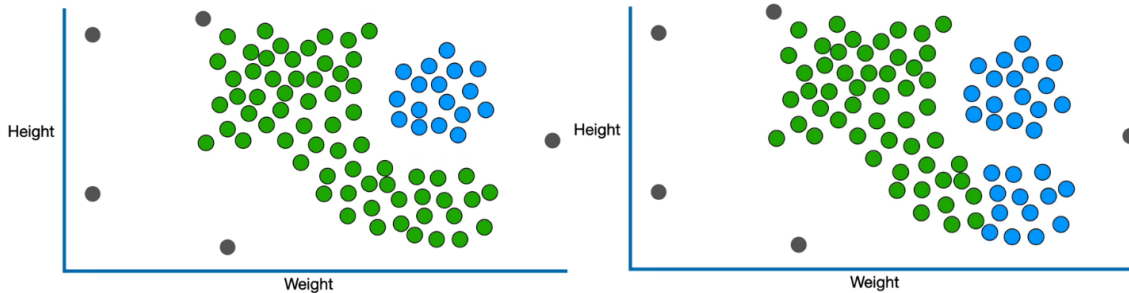
DBSCAN (*Density Based Spatial Clustering of Applications With Noise*) é um algoritmo utilizado para encontrar agrupamentos de diferentes formatos e ruído nas bases de dados, baseado na densidade de objetos no espaço. Galera, nós vimos vários exemplos de pontos plotados em um plano cartesiano e como os algoritmos de agrupamento funcionam sobre eles. Ocorre que alguns métodos funcionam melhor ou pior dependendo do formato dos dados. *Como assim, Diego?*



Na imagem acima, temos um mesmo conjunto de dados agrupados por meio de diversos algoritmos de agrupamento diferentes. Vejam que os resultados variam bastante dependendo da disposição dos pontos de dados. Vocês devem se lembrar também que eu falei que algoritmos como o k-Means são bastante sensíveis a ruídos e anomalias. Além disso, nem sempre temos apenas duas dimensões como na imagem anterior – por vezes, temos dezenas de dimensões.

Vejam o conjunto de dados da imagem à esquerda! Nós, humanos, conseguimos bater o olho e ver com certa facilidade dois grupos e alguns dados anômalos (conforme a imagem à direita).





No entanto, por conta de ruídos, anomalias e do formato dos pontos verdes contornando os pontos azuis, o algoritmo k-Médias teria dificuldade de agrupar os dados podendo gerar algo assim:

Então vamos listar os problemas dos algoritmos vistos até agora: eles têm dificuldades com alguns formatos de pontos de dados; são bastante sensíveis a ruídos e anomalias de dados; e não são ideais para problemas com mais de duas dimensões. É esse o contexto em que o DBSCAN chega para bilhar! Ele é especialista justamente em realizar agrupamentos de dados espaciais (> 2 dimensões) em aplicações com ruído e dados anômalos.

Para tal, ele se baseia na densidade dos pontos de dados, isto é, a quantidade de objetos dentro de um raio de vizinhança. *Como assim, Diego?* Na primeira imagem acima (dos pontos cinzas), como nós – humanos conseguimos identificar dois grupos por meio da densidade dos pontos. Tem um bolinho de pontos juntos de um lado e um bolinha de pontos juntos do outro. O DBSCAN faz algo muito parecido e, por isso, diz-se que ele é baseado em densidade.

Justamente por conta de ser baseado em densidade é que ele lida bem com outliers (dados anômalos). Como a densidade de pontos é baixa próximo de dados anômalos, ele consegue identificá-los com maior facilidade. Não vamos entrar em detalhes do funcionamento do algoritmo em si, basta saber essas características mencionadas acima e que ele funciona buscando a densidade de uma vizinhança dos pontos de dados.

Single/Complete-Linkage

INCIDÊNCIA EM PROVA: baixíssima

Para finalizar os algoritmos de agrupamento, vamos falar bem rapidamente sobre o Single-Linkage e o Complete-Linkage. O Single-Linkage é um método aglomerativo de agrupamento hierárquico no qual novos grupos são criados unindo os grupos mais semelhantes. O agrupamento inicial é formado apenas por singletons (grupos formados por apenas um objeto), e a cada iteração um novo grupo é formado por meio da união dos dois grupos mais similares da iteração anterior.

Neste método, a distância (proximidade) entre o novo grupo e os demais é determinada como a menor distância entre os elementos do novo grupo e os grupos remanescentes. Já o complete-linkage é o inverso: a distância do novo grupo aos demais é calculada como a distância máxima

entre os elementos do novo grupo aos grupos restantes. Eu ainda não vi caindo em provas de tecnologia da informação – apenas em provas de estatístico cujo foco é diferente do nosso.

Regras de Associação

INCIDÊNCIA EM PROVA: ALTA

Uma das principais tecnologias de mineração de dados envolve a descoberta de regras de associação. O banco de dados é considerado uma coleção de transações, cada uma envolvendo um conjunto de itens. Um exemplo comum é o de dados de um carrinho de supermercado. Nesse contexto, o carrinho de supermercado corresponde aos conjuntos de itens que um consumidor compra em um supermercado durante uma visita.

Para esse exemplo, podemos definir as seguintes regras de associação: se leite for comprado, açúcar provavelmente também será; se açúcar for comprado, leite provavelmente também será; se leite e açúcar forem comprados, café provavelmente também será em 60% das transações. **Percebam que geralmente as regras de associação são escritas em um formato como: se [algo acontecer], então [algo acontecerá] ou se [evento], então [ações].**

TRANSACTION	ITEM1	ITEM2	ITEM3
1			
2			
3			
4			
5			

É possível pensar em dezenas de outros exemplos! **Mulheres que compram um sapato em uma loja tendem a comprar uma bolsa também na mesma loja.** Músicos que compram uma nova guitarra tendem a comprar também novas palhetas. *Vocês se lembram do exemplo das cervejas e das fraldas?* Pois é, também são outro excelente exemplo de como utilizar regras de associação para minerar dados. Vamos resumir...

Na mineração de dados, uma regra de associação é um evento que relaciona a presença de um conjunto de itens com outra faixa de valores de um outro conjunto de variáveis. **Uma regra de associação pode ser vista como uma expressão da forma $X \rightarrow Y$, onde há a relação dos**

valores de X e Y em um certo conjunto de valores (Ex: {fralda} → {cerveja}). Isso significa que quem compra leite e fralda também tende a comprar cerveja na mesma transação²⁰ [25].

Existem duas variações comuns de regras de associação: padrões sequenciais e os padrões temporais. Nos padrões sequenciais, uma sequência de ações é buscada. Exemplo: se um paciente passou por uma cirurgia de ponte de safena para artérias bloqueadas e um aneurisma e, depois, desenvolveu ureia sanguínea alta dentro de um ano de cirurgia, ele provavelmente sofrerá de insuficiência renal nos próximos dezoito meses.

Já nos padrões temporais (ou padrões dentro de série temporal), as similaridades podem ser detectadas **dentro de posições de uma série temporal de dados**, que é uma sequência de dados tomados em intervalos regulares, como vendas diárias ou preços de ações de fechamento diário.

Exemplo: uma ação de uma companhia de energia e outra de uma companhia financeira tiveram o mesmo padrão durante um período de três anos em relação a preço de fechamento de ações.

Vamos diferenciar tudo isso? **A técnica de regras de associação visa simplesmente descobrir o relacionamento ou correlação entre variáveis de um banco de dados.** Já a técnica de Padrões Sequenciais busca descobrir padrões sequenciais de eventos de forma equivalente a certos relacionamentos temporais. Por fim, a técnica de Padrões Temporais é bastante semelhante à técnica de Padrões Sequenciais, mas sempre envolve um fator temporal que permite diferenciá-los.

Existem duas medidas capazes de indicar a qualidade ou grau de certeza de uma regra de associação. São elas: suporte e confiança. Vejamos...

MEDIDAS DE INTERESSE	DESCRIÇÃO
SUPORTE/ Prevalência	Trata-se da <u>frequência</u> com que um conjunto de itens específicos ocorrem no banco de dados, isto é, o percentual de transações que contém todos os itens em um conjunto. Em termos matemáticos, a medida de suporte para uma regra $X \rightarrow Y$ é a frequência em que o conjunto de itens aparece nas transações do banco de dados. Um suporte alto nos leva a crer que os itens do conjunto X e Y costumam ser comprados juntos, pois ocorrem com alta frequência no banco (Ex: 70% das compras realizadas em um mercado contém arroz e refrigerante).
CONFIANÇA/ Força	Trata-se da <u>probabilidade</u> de que exista uma relação entre itens. Em termos matemáticos, a medida de confiança para uma regra $X \rightarrow Y$ é a força com que essa regra funciona. Ela é calculada pela frequência dos itens Y serem comprados dado que os itens X foram comprados. Uma confiança alta nos

leva a crer que exista uma alta probabilidade de que se X for comprado, Y também será (Ex: existe uma probabilidade de 70% de que clientes que compram fraldas também comprem cerveja).

Vamos ver um exemplo de cálculo de suporte e confiança porque esse conteúdo será importante para entender o funcionamento dos algoritmos de regras de associação. Vejamos a tabela a seguir:

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

Regra de Associação: {maçã Cerveja}

Para realizar a medida de suporte, basta calcular a razão da sua frequência pela quantidade de transações. Como maçã aparece em 4 das 8 transações, possui um suporte de 50%:

$$\text{Support } \{\text{🍏}\} = \frac{4}{8}$$

Para realizar a medida de confiança, basta calcular em quantas dessas quatro transações ocorreu cerveja como consequente. Como cerveja aparece em 3 das 4 transações, temos

confiança de 75%.

$$\text{Confidence} \{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support} \{\text{🍎, 🍺}\}}{\text{Support} \{\text{🍎}\}} = \frac{3}{4}$$

Bacana! Agora estamos prontos para entender os principais algoritmos de regras de associação. Vamos iniciar pelo algoritmo *apriori*.

Apriori

O algoritmo *apriori* é um método de mineração de dados não supervisionado utilizado para minerar conjuntos de dados frequentes e regras de associação relevantes. Para entender seu funcionamento, vamos ver um exemplo com dados hipotéticos e vamos passar com cada um dos passos quem permitem chegar a essas regras. A tabela seguinte apresenta um conjunto de transações com os respectivos produtos comprados em um supermercado hipotético:

Transação	Produtos
1	Cerveja, Vinho, Queijo
2	Cerveja, Batata Chips
3	Ovos, Farinha de Trigo, Manteiga, Queijo
4	Ovos, Farinha de Trigo, Manteiga, Cerveja, Batata Chips
5	Vinho, Queijo
6	Batata Chips
7	Ovos, Farinha de Trigo, Manteiga, Vinho, Queijo
8	Ovos, Farinha de Trigo, Manteiga, Cerveja, Batata Chips
9	Vinho, Cerveja
10	Cerveja, Batata Chips
11	Manteiga, Ovos
12	Cerveja, Batata Chips

13	Farinha de Trigo, Ovos
14	Cerveja, Batata Chips
15	Ovos, Farinha de Trigo, Manteiga, Vinho, Queijo
16	Cerveja, Vinho, Batata Chips, Queijo
17	Vinho, Queijo
18	Cerveja, Batata Chips
19	Vinho, Queijo
20	Cerveja, Batata Chips

O primeiro passo do algoritmo é calcular a medida de suporte para cada item individual. *Por que?* Porque precisamos reduzir o custo computacional! A quantidade de regras de associação possíveis de serem geradas em bases transacionais cresce exponencialmente com o número de itens da base. Um conjunto de dados simples e pequeno já é capaz de produzir centenas de regras de associação, onerando o processamento dos dados.

Qualquer algoritmo de força bruta, ou seja, que gera todas as combinações possíveis de itens para depois buscar as regras relevantes, não será computacionalmente factível para diversas bases transacionais. Dito isso, uma forma de reduzir o custo computacional dos algoritmos de mineração de regras de associação é desacoplar os requisitos de suporte e os requisitos de confiança mínima das regras.

Como o suporte da regra depende apenas do conjunto de itens, conjuntos de itens pouco frequentes podem ser eliminados no início do processo sem que seja necessário calcular sua confiança. Eu sei que deve ter embaralhado a cabeça de vocês, então eu vou explicar um pouco melhor! *Qual é o objetivo do algoritmo apriori?* Minerar conjuntos de itens frequentes a partir de regras de associação relevantes: com alto suporte a alta confiança.

Uma maneira de fazer isso seria ir testando todas as combinações possíveis de itens até encontrar regras de associação significativas, mas isso tem um custo computacional tão alto que torna essa estratégia inviável. Dito isso, antes de me preocupar em calcular a confiança, eu posso inicialmente calcular o suporte. *Por que?* Porque se o suporte for baixo, eu já descarto esse item! *Isso faz sentido para vocês?* Se um item tem um suporte baixo, significa que ele tem uma frequência baixa.

No contexto de um supermercado, um item com suporte baixo pode ser jiló. Quase ninguém gosta de jiló, logo ele deve aparecer em pouquíssimas transações, portanto possui um baixo suporte. *Entendido?* Então vamos voltar para o nosso exemplo! Eu preciso calcular o suporte

para cada item individual. Como o suporte é apenas a frequência (número de ocorrências) de um determinado produto, basta calcular as quantidades:

Produto	frequência
Vinho	8
Queijo	8
Cerveja	11
Batata chips	10
Ovos	7
Farinha de trigo	6
Manteiga	6

O segundo passo do nosso algoritmo é decidir qual será o suporte mínimo! Produtos cujo suporte seja menor que esse limite mínimo são produtos com baixa frequência que podem ser descartados. *Como esse patamar é escolhido?* Ele é escolhido basicamente de forma arbitrária! Em nosso caso, vamos escolher um suporte mínimo de 7. Isso significa que produtos com suporte (frequência de ocorrência) abaixo desse valor serão descartados: Farinha de Trigo e Manteiga.

O terceiro passo consiste em realizar o mesmo procedimento anterior, mas agora utilizando pares de produtos em vez de produtos individuais. Como já descartamos alguns produtos no passo anterior, teremos um pouco menos de combinações para testar. Nós vamos ignorar todas as combinações que contêm Farinha de Trigo e Manteiga. O resultado dessa análise é apresentado na tabela seguinte:

Conjunto de itens	frequência
Vinho, queijo	7
Vinho, cerveja	2
Vinho, batata chips	1

Vinho, ovos	2
Queijo, cerveja	2
Queijo, batata chips	1
Queijo, ovos	3
Cerveja, batata chips	9
Cerveja, ovos	2
Batata chips, ovos	2

Da mesma forma, é possível notar que apenas duas combinações possuem medida de suporte maior que 7: {Vinho, Queijo} e {Cerveja, Batata Chips}. O próximo passo é continuar com o procedimento anterior, mas para conjuntos maiores. Já testamos produtos individuais, testamos conjuntos de dois itens e agora é o momento de testar conjuntos de três itens. Vamos ver todas as possíveis combinações:

Conjunto de itens	frequência
Vinho, queijo, cerveja	1
Vinho, QUEIJO, batata chips	1
Cerveja, batata chips, vinho	1
Queijo, cerveja, batata chips	1

Note que nenhum conjunto de três itens contém um suporte maior que o mínimo pré-definido. Na verdade, nem era necessário calcular – bastava lembrar que {Queijo, Cerveja}, {Queijo, Batata Chips}, {Vinho, Batata Chips} e {Queijo, Cerveja} já haviam sido descartados no passo anterior. Agora que já temos o maior conjunto de itens frequentes, basta gerar as regras de associação e realizar o cálculo da confiança. *Fechado?*

Bem, como todos os conjuntos de três itens foram descartados, então ficamos com os conjuntos de dois itens: {Vinho, Queijo} e {Cerveja, Batata Chips}. Agora vamos pegar todos os subconjuntos desses conjuntos, criar as regras e verificar se elas possuem um patamar mínimo

de confiança. *E qual seria esse valor, Diego?* Da mesma forma do suporte, trata-se de um valor arbitrário – nós vamos escolher o valor mínimo de 85%. Para cada subconjunto S de I, teremos a regra:

$$S \rightarrow (I - S)$$

Logo, se temos um conjunto de dados $I = \{A, B, C\}$, temos os subconjuntos $\{A\}$, $\{B\}$, $\{C\}$, $\{A,B\}$, $\{A, C\}$ e $\{B,C\}$. Dessa forma, uma regra poderia ser: $\{A\} \rightarrow \{A, B, C\} - \{A\}$, logo $\{A\} \rightarrow \{B, C\}$. Vamos lá...

- **Conjunto de Itens 1:** {Vinho, Queijo}
- **Subconjuntos:** {Vinho}, {Queijo}, {Vinho, Queijo}

Regra de Associação 1:

$$\{Vinho\} \rightarrow \{Vinho, Queijo\} - \{Vinho\} = \{Vinho\} \rightarrow \{Queijo\}$$

Regra de Associação 2:

$$\{Queijo\} \rightarrow \{Vinho, Queijo\} - \{Queijo\} = \{Queijo\} \rightarrow \{Vinho\}$$

- **Conjunto de Itens 2:** {Cerveja, Batata Chips}
- **Subconjuntos:** {Cerveja}, {Batata Chips}, {Cerveja, Batata Chips}

Regra de Associação 3:

$$\{Cerveja\} \rightarrow \{Cerveja, Batata Chips\} - \{Cerveja\} = \{Cerveja\} \rightarrow \{Batata Chips\}$$

Regra de Associação 4:

$$\{Batata Chips\} \rightarrow \{Cerveja, Batata Chips\} - \{Batata Chips\} = \{Batata Chips\} \rightarrow \{Cerveja\}$$

Lembrando que desconsideramos regras cujo antecedente ou consequente sejam nulos. Agora é o momento de calcular a Confiança = Suporte(I)/Suporte(S). Nós já calculamos esses valores antes:

Produto	frequência
Vinho	8
Queijo	8

Cerveja	11
Batata chips	10

Conjunto de itens	frequência
Vinho, queijo	7
Cerveja, batata chips	9

O cálculo da confiança é bastante intuitivo: nós vamos ver a proporção – dentre as transações que contenham o conjunto de itens – daquelas transações em que a regra de transação é válida. Por exemplo: dada uma regra de associação $X \rightarrow Y$, se temos 10 transações em que ocorre o item $\{X\}$, vamos calcular em quantas dessas transações X ocorreu também Y . *Entendido?* Então, faremos isso para as quatro regras de transação descobertas:

Regra de Associação 1: $\{\text{Vinho}\} \rightarrow \{\text{Queijo}\}$

Confiança: $\text{Suporte}(\text{Vinho}, \text{Queijo}) / \text{Suporte}(\text{Vinho}) = 7/8 = 87,5\% > 85\%$ (**Regra aceita**)

Regra de Associação 2: $\{\text{Queijo}\} \rightarrow \{\text{Vinho}\}$

Confiança: $\text{Suporte}(\text{Queijo}, \text{Vinho}) / \text{Suporte}(\text{Queijo}) = 7/8 = 87,5\% > 85\%$ (**Regra aceita**)

Regra de Associação 3: $\{\text{Cerveja}\} \rightarrow \{\text{Batata Chips}\}$

Confiança: $\text{Suporte}(\text{Cerveja}, \text{Batata Chips}) / \text{Suporte}(\text{Cerveja}) = 9/11 = 81\% < 85\%$ (**Regra rejeitada**)

Regra de Associação 4: $\{\text{Batata Chips}\} \rightarrow \{\text{Cerveja}\}$

Confiança: $\text{Suporte}(\text{Batata Chips}, \text{Cerveja}) / \text{Suporte}(\text{Batata Chips}) = 9/10 = 90\% > 85\%$ (**Regra aceita**)

Em síntese: o algoritmo *apriori* é um método baseado no conceito de itens frequentes que constrói um conjunto de regras das mais simples (único item) às mais complexas (múltiplos itens), sendo que – para cada nível de regra – calcula-se o número de ocorrências nos dados (suporte) e eliminam-se as regras com suporte inferior a um determinado patamar mínimo de tal forma que regras que subsistirem sejam expandidas para mais um produto e assim por diante.

FP-Growth

O FP-Growth (*Frequent Pattern – Growth*) é uma técnica utilizada para descobrir padrões de associação entre itens em um conjunto de dados muito eficaz para trabalhar com grandes conjuntos de dados usando uma memória limitada. A técnica é baseada na construção de uma árvore de frequência de itens (em uma abordagem bottom-up), que é usada para descobrir padrões frequentes nos dados.

Inicia-se com os itens individuais e eles são agrupados em padrões mais complexos. O algoritmo é muito eficiente e pode ser usado para descobrir padrões em conjuntos de dados muito grandes. O método *apriori* tem dificuldade para tratar uma grande quantidade de conjuntos candidatos e execução de repetidas passagens pela base de dados. Já a árvore é capaz de armazenar de forma comprimida a informação sobre padrões frequentes. Esse método ainda não foi cobrado em prova.

CRISP-DM

INCIDÊNCIA EM PROVA: baixa

Galera, vocês já devem ter notado que nós – malucos da área de tecnologia da informação – adoramos processos! É impressionante, nem advogados e juízes gostam tanto de processos! Vejam só: nós temos um modelo de referência para gerenciamento de projetos, um para análise de negócio, um para gerenciamento de serviços, um para qualidade de software, um para governança de tecnologia da informação... **e todos eles são baseados em processos!**

Ora, se nós temos modelos de referência para esse bando de coisas, por que não temos um modelo de referência para a mineração de dados? Pois é, mas nós já temos!!! O **CRISP-DM** (**C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) é um modelo de referência²¹_[26] de mineração de dados que descreve um conjunto de processos para realizar projetos de mineração de dados em uma organização baseado nas melhores práticas utilizadas por profissionais e acadêmicos do ramo.

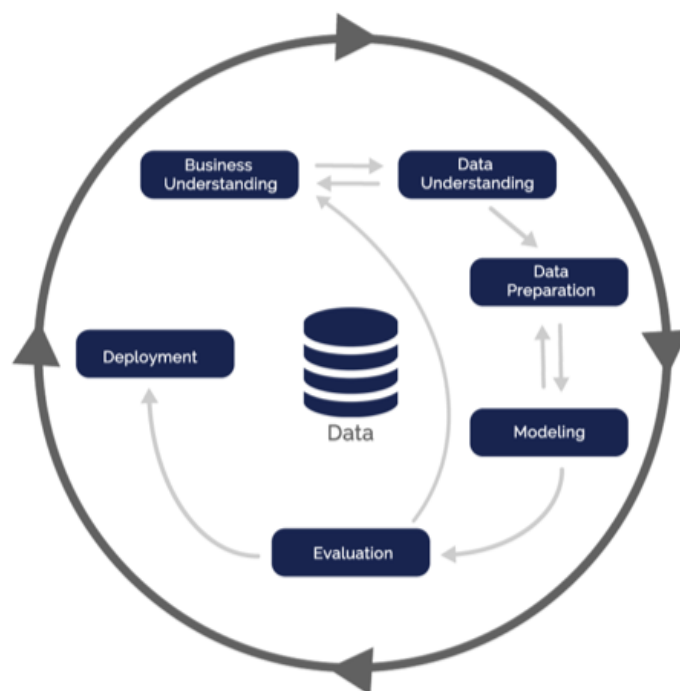
Galera, como nascem todos esses modelos de referência? **Em geral, reúnem-se os maiores especialistas da área e eles exibem como fazem para resolver problemas recorrentes.** As ideias são, então, organizadas em forma de processos e tarefas em um documento de modo que outras pessoas que desejem resolver problemas semelhantes (em nosso caso, projetos de mineração de dados) possam usá-lo como referência. *Entendido?* Vamos ver agora o que dizia o site oficial:

“O Projeto CRISP-DM desenvolveu um modelo de processos de mineração de dados com foco industrial e independente de ferramentas. Partindo dos processos embrionários de descoberta de conhecimento usados atualmente na indústria e respondendo diretamente aos requisitos do usuário, este projeto definiu e validou um processo de mineração de dados aplicável em diversos setores da indústria. Isso tornará grandes projetos de mineração de dados mais rápidos, mais baratos, mais confiáveis e mais gerenciáveis. Até casos de mineração de dados em pequena escala se beneficiarão do uso do CRISP-DM”.

Dessa forma, caso você queira fazer um projeto de mineração de dados em sua organização, você poderá utilizar esse modelo de processos como referência – **é importante destacar que se trata de uma metodologia não proprietária que pode ser aplicada livremente a qualquer projeto independentemente do tamanho ou tipo do negócio.** Bem, essa metodologia possui um ciclo de vida não-linear composto por seis fases ou etapas. Vejamos...

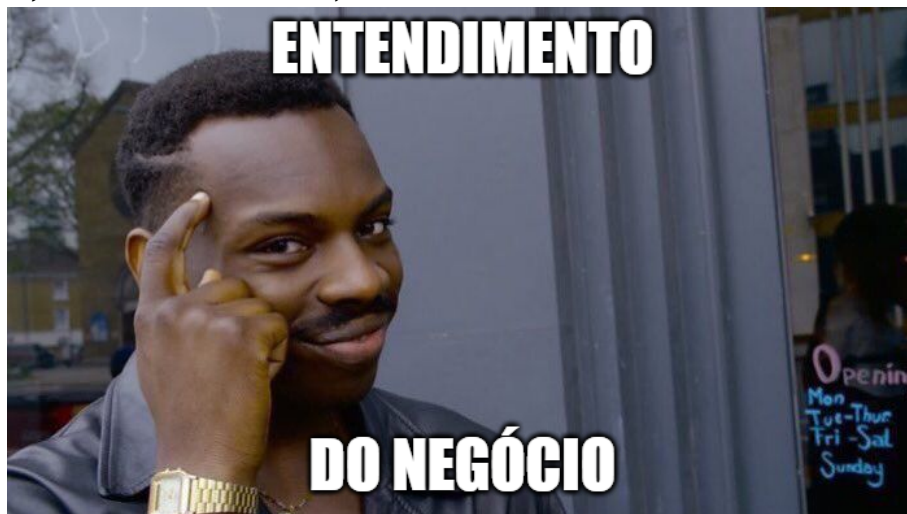
*Professor, essas fases devem ser executadas em uma sequência rigorosa? Não, conforme é possível ver na imagem a seguir, é sempre necessário ir e voltar entre diferentes fases – e isso depende do resultado de cada fase. **Observem também na imagem que as setas indicam as dependências mais importantes e frequentes entre as fases.*** Além disso, o círculo externo simboliza a natureza cíclica da própria mineração de dados. *Como assim, Diego?*

Um processo de mineração de dados continua após a implantação de uma solução. As lições aprendidas durante o processo podem desencadear novas questões comerciais, geralmente mais focadas. Os processos subsequentes de mineração de dados se beneficiarão das experiências dos anteriores. *Beleza? As fases são: (1) Entendimento do Negócio; (2) Entendimento dos Dados; (3) Preparação dos Dados; (4) Modelagem; (5) Avaliação; e (6) Implantação.*



Entendimento do Negócio

Essa fase inicial concentra-se no entendimento dos objetivos e requisitos do projeto de uma perspectiva de negócio e, em seguida, na conversão desse conhecimento em uma definição de problema de mineração de dados e em um plano preliminar desenvolvido para atingir os objetivos. **Em outras palavras, essa fase busca entender qual problema o negócio quer resolver!** *Professor, não entendi! Calma, vamos lá...*



Pessoal, é muito comum que uma área de tecnologia da informação faça um projeto de mineração de dados para uma área que ela não domina o assunto! Você pode aplicar a tecnologia de mineração à área de saúde, finanças, turismo, esportes, comércio, etc. **A galera da área de tecnologia entende de tecnologia, não entende de finanças por exemplo. Logo, antes de começar o projeto, é importantíssimo que ela entenda do negócio.**

Essa é a fase em que os analistas de tecnologia vão entender qual é o problema que o negócio quer resolver, seus objetivos, como está a situação atual, quais são os requisitos do projeto, quais são os principais pressupostos e limitações, em qual forma será a entrega dos resultados, quais são os critérios de sucesso, entre outros parâmetros – **tudo isso para desenvolver um planejamento de projeto.** *Bacana?*

Entendimento dos Dados

A segunda fase começa com uma coleta inicial dos dados e prossegue com atividades para explorá-los com o intuito de obter um maior conhecimento e familiaridade. **Em seguida, busca-se avaliar a qualidade dos dados, descobrir as primeiras ideias sobre os dados ou detectar subconjuntos interessantes para formar hipóteses de informação ocultas e descobrir insights.** Essa fase também é responsável por descrever os dados – por vezes, utilizando estatísticas.



Na descrição, pode-se obter uma espécie de fotografia dos dados contendo a localização, o formato, a fonte, o número de registros, o número de atributos, como será a extração dos dados, e outras características que interessem, assegurando que esses dados consigam representar o problema em análise. **Esta etapa também envolve o que geralmente é denominado análise exploratória de dados.** Enfim... o lance aqui é ser o mais íntimo possível dos dados em análise.

Preparação dos Dados



Também chamada de pré-processamento, nessa fase ocorre a preparação dos dados para a fase de modelagem. Essa etapa ocorre quando já entendemos o problema do negócio e já exploramos os dados disponíveis. Ela abrange todas as atividades para construir o conjunto de dados final a partir dos dados brutos iniciais, isto é, aqueles que serão alimentados na ferramenta de modelagem. *Professor, e que atividades são essas?* Bem, me acompanhem...

Essa lista não é exaustiva, mas inclui tarefas como seleção de tabelas, integração, transformação, limpeza e organização de dados – além da seleção e engenharia de recursos. Essas atividades visam a melhoria na qualidade dos dados originais e para realizá-las existem ferramentas de mineração de dados que dispõem de funcionalidades específicas que garantem agilidade nas operações, ou seja, você não precisa fazer isso “na mão”.

Galera, é bem provável que as tarefas de preparação de dados sejam executadas várias vezes, de forma iterativa e sem nenhuma sequência predefinida. **Além disso, trata-se da fase mais demorada, ocupando mais de 70% do tempo/esforço total gasto em qualquer projeto de**

ciência de dados. *Por que?* Porque ela é a responsável por carregar os dados identificados na etapa anterior e prepará-los para análise por meio de métodos de mineração de dados.

Vamos falar em uma linguagem mais clara agora: no mundo real, os dados não estão todos organizados bonitinhos prontos para serem analisados. **Tem dado faltando, dado incompleto, dado errado, dados discrepantes, dados inconsistentes, dados em formatos estranhos, entre outros.** *Como nós vamos analisar dados totalmente desestruturados?* Logo, essa etapa busca organizá-los da melhor maneira possível.

Dessa forma, é necessário integrar os dados! **Há momentos em que os dados estão disponíveis em várias fontes e, portanto, precisam ser combinados com base em determinadas chaves ou atributos para melhor uso.** É necessário também selecionar os dados desejados no modelo! Por exemplo: talvez você não queira usar dados *outliers* (fora da curva) ou talvez você não queira utilizar todas as colunas de uma tabela. Enfim, escolha tudo que será relevante para seu modelo!

Legal, mas o que fazemos em relação aos dados incorretos? Bem, nós devemos limpá-los! **Dados em formatos incorretos e números inteiros sendo interpretados como textos são exemplos de “sujeira” que podem ser encontradas no seu dado – esse é o momento de tratá-las.** Há também momentos em que precisamos derivar ou gerar recursos a partir dos existentes. *Como assim?* Ex: algumas vezes, você deve derivar a idade de uma pessoa a partir da data de nascimento.

Enfim... todas essas atividades são realizadas para tratar a qualidade dos dados da melhor forma possível de modo que possam ser utilizadas pelos algoritmos de mineração de dados. *Fechou?*

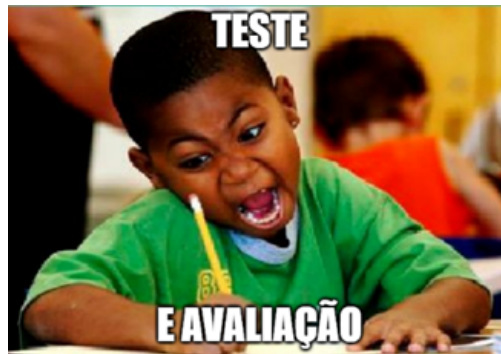
Construção do Modelo



Também chamada de Modelagem, nessa fase ocorre a seleção das técnicas, ferramentas e algoritmos a serem utilizados, como também a elaboração e execução da modelagem sobre o conjunto de dados preparado na fase anterior. Aliás, retornar à fase de preparação é bem frequente e necessário nessa etapa – lembrem-se que se trata de um processo iterativo e cheio de vai e volta! Você pode criar diferentes modelos e compará-los na próxima fase! *Bacana?*

Em outras palavras, essa etapa utiliza os dados limpos e formatados preparados na etapa anterior para fins de modelagem. **Ela inclui a criação, avaliação e ajuste fino de modelos e parâmetros para valores ideais, com base nas expectativas e critérios estabelecidos durante a fase de entendimento dos negócios.** Dependendo da necessidade do negócio, a tarefa de mineração de dados pode ser de uma classificação, uma regressão, uma associação, uma clusterização, etc.

Teste e Avaliação



Bem, chegou a hora de avaliar os resultados da modelagem. *Você se lembra que nós definimos critérios de sucesso lá na primeira fase?* **Agora é a hora de verificar se ela foi atingida. Se não foi, é necessário voltar a primeira fase e entender o que deu de errado, determinar um novo escopo e tentar novamente.** Caso tudo tenha dado certo, é hora de seguir em frente para a última fase. *E o que mais?*

Nessa fase do projeto, você construiu um modelo (ou vários modelos) que parece ter alta qualidade do ponto de vista da análise de dados. Antes de prosseguir para a implantação final do modelo, é importante avaliá-lo mais detalhadamente e revisar as etapas executadas para garantir que ele atinja adequadamente os objetivos de negócios. No final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada.

Implantação/Implementação



Também chamada de desenvolvimento, essa fase busca colocar o modelo para funcionar. Ele coloca fim ao seu projeto, mas é necessário se lembrar de monitorar os resultados e de adaptar o modelo sempre que necessário. Os modelos que foram desenvolvidos, ajustados, validados e testados durante várias iterações são salvos e preparados para o ambiente de produção (o nome é estranho, mas esse é o ambiente em que o software está de fato funcionando).

É necessário criar um plano de implantação adequado que inclui detalhes sobre os requisitos de hardware e software. **O estágio de implantação também inclui a verificação e o monitoramento de aspectos para avaliar o modelo em produção quanto a resultados, desempenho e outras métricas.** Dependendo dos requisitos, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados repetível.

Em muitos casos, será o cliente, não o analista de dados, quem executará as etapas de implantação. **No entanto, mesmo que o analista não realize o esforço de implantação, é importante que o cliente entenda antecipadamente quais ações precisarão ser executadas para realmente fazer uso dos modelos criados.** Entendido, galera? Agora vamos fazer um resumo de tudo para ver se vocês realmente entenderam...

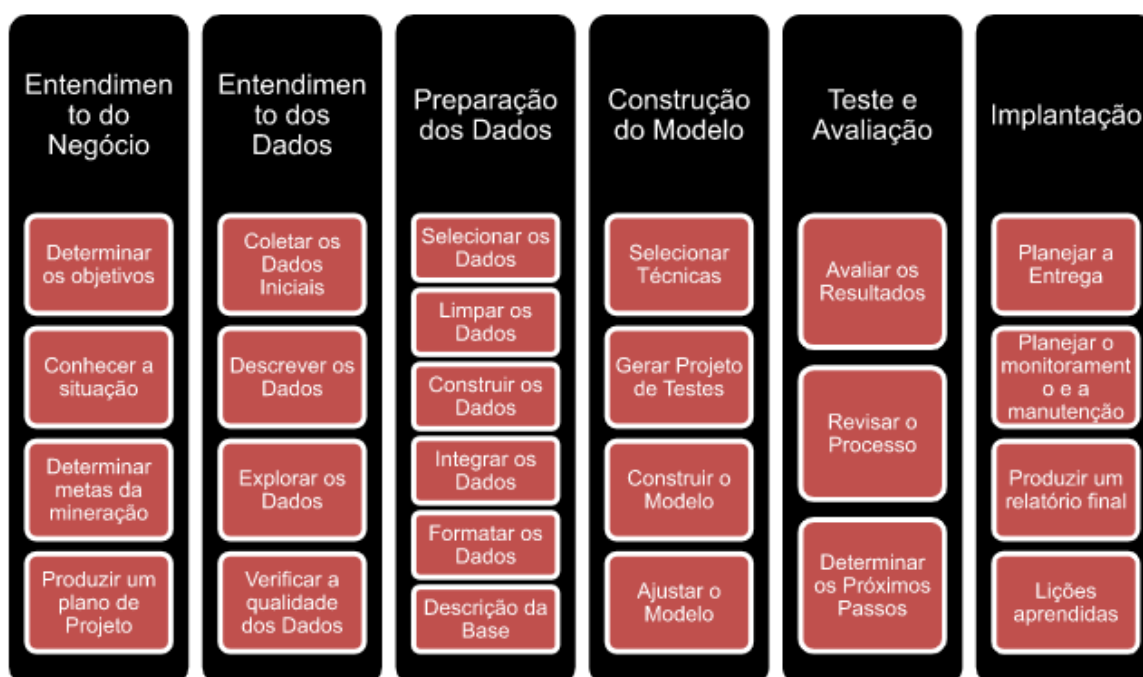
Acompanhem o Tio Diego aqui! Vamos supor que eu tenha sido chamado para fazer um projeto de mineração de dados da Bolsa de Valores! *Eu entendo algo desse negócio?* Não, então meu primeiro passo é entender o negócio. **Eu vou lá na Bolsa de Valores, entrevisto algumas pessoas, converso com outras até entender qual é o problema que se quer resolver, quais são os objetivos, os requisitos, entre outros.** *Certinho até aqui?* Bacana...

Em seguida, eu vou lá ver como estão os dados que serão utilizados. Eu vejo que os dados vêm de dez sistemas diferentes, analiso como está a qualidade desses dados, qual é a quantidade, entre outras coisas. *Para quê?* **Para que eu me familiarize com os dados! Em seguida, eu vou pré-processar os dados.** Ora, tem dado corrompido, inconsistente, incompleto, faltante, etc – eu preciso fazer aquela limpeza básica para começar a trabalhar. Vamos para a modelagem...

Agora vou escolher ferramentas, técnicas e algoritmos que serão utilizados para modelar os meus dados. *Que técnicas eu posso utilizar?* Eu posso utilizar a classificação, a estimativa, a previsão, a análise de afinidades, a análise de agrupamentos, entre outras. *E que algoritmos?* Eu

posso utilizar, por exemplo, árvores de decisão ou redes neurais. *E que ferramentas?* Eu posso utilizar *SAS Enterprise Miner* ou *IBM Intelligent Miner* ou *Oracle Darwin Data Mining Software*.

Em seguida, eu vou testar e avaliar os modelos desenvolvidos quanto à precisão e generalidade. *Foram atendidos os objetivos de negócio?* **Se sim, partimos para a fase de implantação, que é colocar o modelo para funcionar!** Segue abaixo uma imagem com as principais atividades de cada fase. Galera, esse não é um tema que cai muito em prova – na verdade, eu só encontrei cinco questões. Saibam disso para dosar bem os estudos de vocês. *Fechou? ;)*



13.2 Análise de Informações Mineração de Dados - Resumo

Resumo

DEFINIÇÕES DE DATA MINING

Data Mining é o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida.

Palavras-chave: exploração; informação implícita desconhecida.

Data Mining é uso de teorias, métodos, processos e tecnologias para organizar uma grande quantidade de dados brutos para identificar padrões de comportamentos em determinados públicos.

Palavras-chave: teorias; métodos; processos; tecnologias; organizar dados brutos; padrões de comportamentos.

Data Mining é a categoria de ferramentas de análise denominada open-end e que permite ao usuário avaliar tendências e padrões não conhecidos entre os dados.

Palavras-chave: ferramenta de análise; open-end; tendências e padrões.

Data Mining é o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

Palavras-chave: descoberta; correlações; padrões; tendências; reconhecimento de padrões; estatística; matemática.

Data Mining constitui em uma técnica para a exploração e análise de dados, visando descobrir padrões e regras, a princípio ocultos, importantes à aplicação.

Palavras-chave: exploração e análise de dados; padrões; regras; ocultos.

Data Mining é o conjunto de ferramentas que permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de computação como redes neurais, algoritmos genéticos e lógica nebulosa (fuzzy), dentre outras.

Palavras-chave: tendências; padrões; redes neurais; algoritmos genéticos; lógica nebulosa.

Data Mining é o conjunto de ferramentas e técnicas de mineração de dados que têm por objetivo buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.

Palavras-chave: classificação; agrupamento; clusterização; padrões.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes com o intuito de detectar relacionamentos sistemáticos entre variáveis e novos subconjuntos de dados.

Palavras-chave: padrões; relacionamentos.

Data Mining consiste em explorar um conjunto de dados visando a extrair ou a ajudar a evidenciar padrões, como regras de associação ou sequências temporais, para detectar relacionamentos entre estes.

Palavras-chave: exploração; padrões; regras; associação; sequência temporal; detecção.

Data Mining são ferramentas que utilizam diversas técnicas de natureza estatística, como a análise de conglomerados (cluster analysis), que tem como objetivo agrupar, em diferentes conjuntos de dados, os elementos identificados como semelhantes entre si, com base nas características analisadas.

Palavras-chave: estatística; análise de conglomerados; agrupamento.

Data Mining é o conjunto de técnicas que, envolvendo métodos matemáticos e estatísticos, algoritmos e princípios de inteligência artificial, tem o objetivo de descobrir relacionamentos significativos entre dados armazenados em repositórios de grandes volumes e concluir sobre padrões de comportamento de clientes de uma organização.

Palavras-chave: métodos matemáticos e estatístico; inteligência artificial; relacionamentos; padrões; comportamentos.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

Palavras-chave: padrões; regras de associação; sequências temporais; relacionamentos.

Data Mining é o processo de identificar, em dados, padrões válidos, novos, potencialmente úteis e, ao final, compreensíveis.

Palavras-chave: padrões; utilidade.

Data Mining é um método computacional que permite extrair informações a partir de grande quantidade de dados.

Palavras-chave: extração.

Data Mining é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais.

Palavras-chave: exploração; padrões consistentes; regras de associação; sequência temporal.

Data Mining é o processo de analisar de maneira semi-automática grandes bancos de dados para encontrar padrões úteis.

Palavras-chave: padrões.**CARACTERÍSTICAS de MINERAÇÃO DE DADOS**

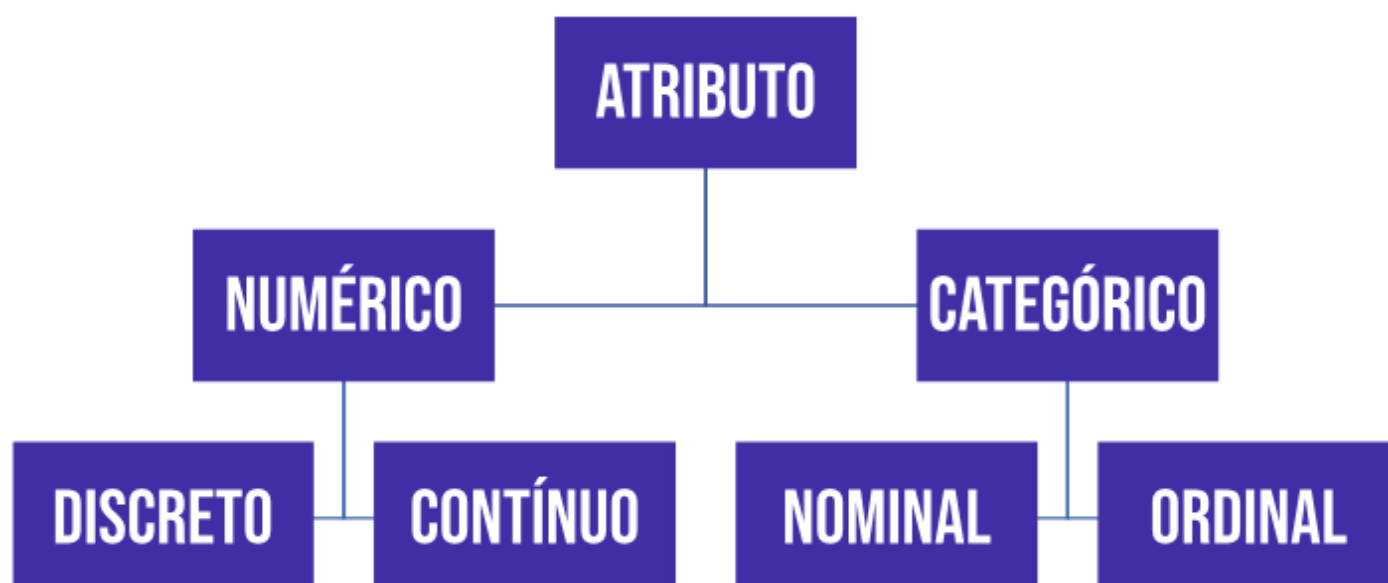
Há diferentes tipos de mineração de dados: (1) diagnóstica, utilizada para entender os dados e/ou encontrar causas de problemas; (2) preditiva, utilizada para antecipar comportamentos futuros.

As provas vão insistir em afirmar que a mineração de dados só pode ocorrer em bancos de dados muito grades como Data Warehouses, mas isso é falso – apesar de comum, não é obrigatório.

Em geral, ferramentas de mineração de dados utilizam uma arquitetura web cliente/servidor, sendo possível realizar inclusive a mineração de dados de bases de dados não estruturadas.

Não é necessário ter conhecimentos de programação para realizar consultas, visto que existem ferramentas especializadas que auxiliam o usuário final de negócio.

TIPOS DE DADOS	DESCRIÇÃO
DADOS ESTRUTURADOS	São aqueles que residem em campos fixos de um arquivo (Ex: tabela, planilha ou banco de dados) e que dependem da criação de um modelo de dados, isto é, uma descrição dos objetos juntamente com as suas propriedades e relações.
DADOS SEMIESTRUTURADOS	São aqueles que não possuem uma estrutura completa de um modelo de dados, mas também não é totalmente desestruturado. Em geral, são utilizados marcadores (<i>tags</i>) para identificar certos elementos dos dados, mas a estrutura não é rígida.
DADOS NÃO ESTRUTURADOS	São aqueles que não possuem um modelo de dados, que não está organizado de uma maneira predefinida ou que não reside em locais definidos. Eles costumam ser de difícil indexação, acesso e análise (Ex: imagens, vídeos, sons, textos livres, etc).
ATRIBUTO DEPENDENTE	Representa um atributo de saída que desejamos manipular em um experimento de dados (também chamado de variável alvo ou variável target).
ATRIBUTO INDEPENDENTE	Representa um atributo de entrada que desejamos registrar ou medir em um experimento de dados.



TIPOS DE ATRIBUTOS	DESCRIÇÃO
ATRIBUTO NUMÉRICO	Também chamado de atributo quantitativo, é aquele que pode ser medido em uma escala quantitativa, ou seja, apresenta valores numéricos que fazem sentido.
discreto	Os valores representam um conjunto finito ou enumerável de números, e que resultam de uma contagem (Ex: número de filhos, número de bactérias por amostra, número de logins em uma página web, entre outros).
contínuo	Os valores pertencem a um intervalo de números reais e representam uma mensuração (Ex: altura de uma pessoa, peso de uma marmita, salário de um servidor público, entre outros).

PROCESSO DE DESCOBERTA DE CONHECIMENTO

Data Mining faz parte de um processo muito maior de descoberta de conhecimento chamada KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Bancos de Dados).

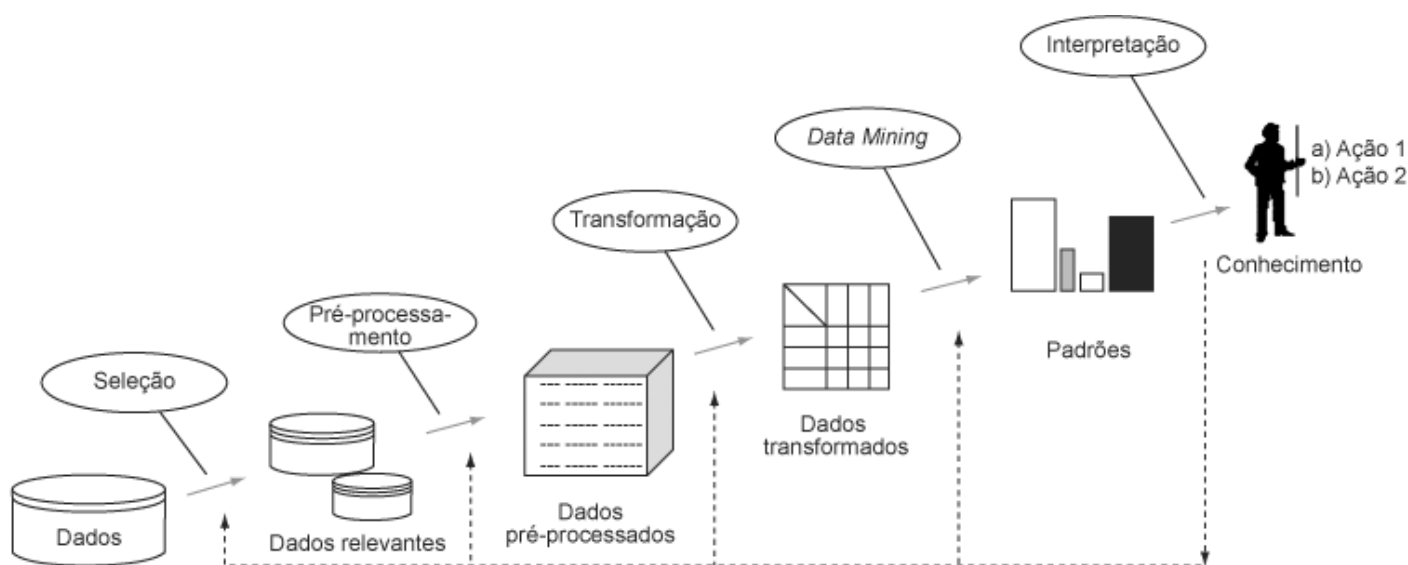


Figura 1. Etapas do processo KDD (Fayyad et al. (1996)).

1	SELEÇÃO DE DADOS	Dados sobre itens ou categorias são selecionados.
---	-------------------------	---

2	LIMPEZA DE DADOS	Dados são corrigidos ou eliminados dados incorretos
3	ENRIQUECIMENTO DE DADOS	Dados são melhorados com fontes de informações adicionais
4	TRANSFORMAÇÃO DE DADOS	Dados são reduzidos por meio de sumarizações, agregações e <u>discretizações</u> . [27]
5	MINERAÇÃO DE DADOS	Padrões úteis são descobertos.
6	EXIBIÇÃO DE DADOS	Informações descobertas são exibidas ou relatórios são construídos.

Objetivos da mineração	DESCRIÇÃO
Previsão	A mineração de dados pode mostrar como certos atributos dos dados se comportarão no futuro. Para realizar a previsão (ou prognóstico), a lógica de negócios é utilizada em conjunto com a mineração de dados (Ex: prever um terremoto com alta probabilidade).
Identificação	Padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade (Ex: padrões de comportamento de hackers permitem identificar possíveis intrusos acessando sistema).
Classificação	A mineração de dados permite particionar os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros (Ex: clientes podem ser categorizados pelos seus perfis de compradores).
otimização	A mineração de dados pode otimizar o uso de recursos limitados, como tempo, espaço, dinheiro ou materiais e maximizar variáveis de saída como vendas ou lucros sob determinadas restrições (Ex: tempo, escopo e custo de um projeto).



Técnicas	DESCRIÇÃO
Classificação	Hierarquia de classes com base em um conjunto existente de eventos ou transações.
Regressão	Aplicação especial da regra de classificação em busca de uma função que mapeie registros de um banco de dados.
Regras de associação	Busca descobrir relacionamentos entre variáveis correlacionando a presença de um item com uma faixa de valores para outro conjunto de variáveis
agrupamento	Particiona dados em segmentos previamente desconhecidos com características semelhantes.

MEDIDAS DE INTERESSE	DESCRIÇÃO
SUPORTE/ Prevalência	Trata-se da <u>frequência</u> com que um conjunto de itens específicos ocorrem no banco de dados, isto é, o percentual de transações que contém todos os itens em um conjunto. Em termos matemáticos, a medida de suporte para

CONFIANÇA/ Força	uma regra $X \rightarrow Y$ é a frequência em que o conjunto de itens aparece nas transações do banco de dados. Um suporte alto nos leva a crer que os itens do conjunto X e Y costumam ser comprados juntos, pois ocorrem com alta frequência no banco (Ex: 70% das compras realizadas em um mercado contém arroz e refrigerante).
	Trata-se da <u>probabilidade</u> de que exista uma relação entre itens. Em termos matemáticos, a medida de confiança para uma regra $X \rightarrow Y$ é a força com que essa regra funciona. Ela é calculada pela frequência dos itens Y serem comprados dado que os itens X foram comprados. Uma confiança alta nos leva a crer que exista uma alta probabilidade de que se X for comprado, Y também será (Ex: existe uma probabilidade de 70% de que clientes que compram fraldas também comprem cerveja).

CONCEITOS AVANÇADOS	DESCRIÇÃO
APRENDIZADO DE MÁQUINA	Trata-se de uma ferramenta poderosa para a aquisição automática de conhecimento por meio da imitação do comportamento de aprendizagem humano com foco em aprender a reconhecer padrões complexos e tomar decisões.
MINERAÇÃO DE TEXTO	Trata-se de um meio para encontrar padrões interessantes/úteis em um contexto de informações textuais não estruturadas, combinado com alguma tecnologia de extração e de recuperação da informação, processo de linguagem natural e de sumarização ou indexação de documentos.

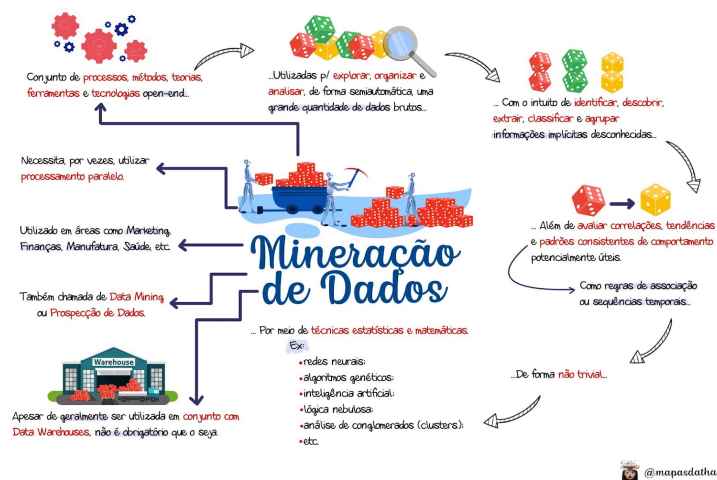


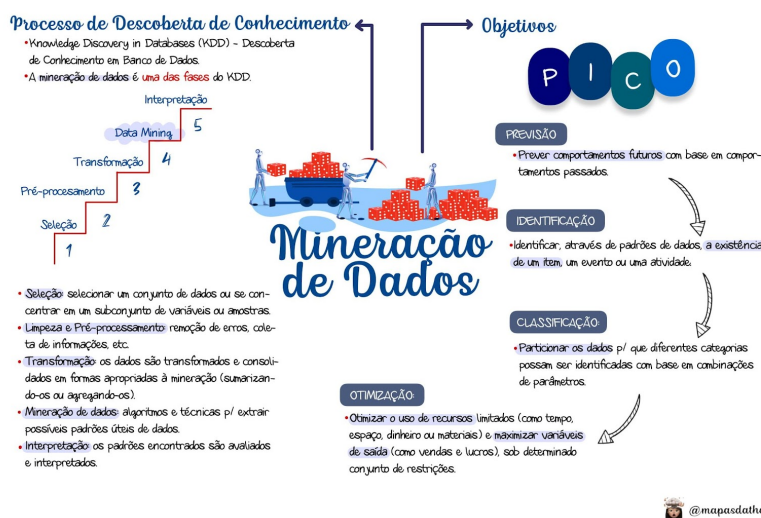
Crisp-dm	DESCRIÇÃO
Entendimento do negócio	Busca compreender das necessidades gerenciais e dos objetivos e requisitos de negócio que devem ser atendidos pela mineração de dados.

Entendimento dos dados	Busca identificar os dados relevantes das diferentes fontes de dados.
Preparação dos dados	Busca carregar os dados identificados no passo anterior e prepará-los para análise por métodos de mineração de dados.
Construção do modelo	Busca selecionar e aplicar técnicas de modelagem a um conjunto de dados previamente preparado.
Teste e avaliação	Busca testar e avaliar os modelos desenvolvidos.
implantação	Busca organizar o conhecimento adquirido com a exploração dos dados de forma que o usuário possa compreendê-lo.

13.3 Mapa Mental Análise de Informações Mineração de Dados

Mapa Mental





13.4 Questões Comentadas Análise de Informações Mineração de Dados Multibancas

#VamosPraticar#

36. (CESPE / ME – 2020)

A análise de regressão em mineração de dados tem como objetivos a sumarização, a predição, o controle e a estimação.

Comentários:

A análise de regressão consiste na realização de uma análise estatística com o objetivo de

verificar a existência de uma relação funcional entre uma variável dependente com uma ou mais variáveis independentes. De maneira geral, a análise de regressão tem como objetivos a sumarização, a predição, o controle, a estimação, a seleção de variáveis e a inferência.

Gabarito: Correto

#VamosPraticar#

#VamosPraticar#

86.(CESPE / Correios – 2011)

Um dos métodos de classificação do datamining é o de análise de agrupamento (cluster), por meio do qual são determinadas características sequenciais utilizando-se dados que dependem do tempo, ou seja, extraíndo-se e registrando-se desvios e tendências no tempo.

Comentários:

A questão se refere ao padrão temporal das regras de associação (e, não, análise de agrupamento), isto é, similaridades podem ser detectadas dentro de posições de uma série temporal de dados, que é uma sequência de dados tomados em intervalos regulares. Seu objetivo é modelar o estado do processo extraíndo e registrando desvios e tendências no tempo.

Gabarito: Errado

87.(CESPE / TJ-ES – 2011)

Mineração de dados, em seu conceito pleno, consiste na realização, de forma manual, de sucessivas consultas ao banco de dados com o objetivo de descobrir padrões úteis, mas não necessariamente novos, para auxílio à tomada de decisão.

Comentários:

Vamos corrigir a redação do item: *Mineração de dados, em seu conceito pleno, consiste na realização, de forma manual* **automática ou semiautomática**, *de sucessivas consultas ao banco de dados com o objetivo de descobrir padrões úteis, mas não necessariamente* **novos**, *para auxílio à tomada de decisão.*

Gabarito: Errado

89.(CESPE / SERPRO – 2010)

A mineração de dados (datamining) é uma atividade de processamento analítico não trivial, que, por isso, deve ser realizada por especialistas em ferramentas de desenvolvimento de software e em repositórios de dados históricos orientados a assunto (datawarehouse).

Comentários:

Em primeiro lugar, a mineração de dados não deve ser realizada em Data Warehouses – apesar de ser comum; em segundo lugar, ela não deve ser realizada por especialistas em ferramentas de desenvolvimento de software (isto é, programadores) e, sim, por usuários finais de negócio.

Gabarito: Errado

91.(CESPE / EMBASA – 2010)

Data mining é o processo de extração de conhecimento de grandes bases de dados, sendo estas convencionais ou não, e que faz uso de técnicas de inteligência artificial.

Comentários:

Eu não vejo absolutamente nenhum erro nessa questão! Ele pode ou não ocorrer em grandes bases de dados, podem ser bases convencionais ou não e pode utilizar técnicas de inteligência artificial ou não, mas a banca a considerou falsa.

Gabarito: Errado

93.(CESPE / IPEA – 2008)

O data mining é um processo utilizado para a extração de dados de grandes repositórios para tomada de decisão, mas sua limitação é não conseguir analisar dados de um data warehouse.

Comentários:

Data Mining não é utilizado para extração de dados e, sim, conhecimentos ou insights de grandes repositórios para tomada de decisão. Além disso, ele não só consegue como frequentemente é utilizado para analisar dados de um Data Warehouse. A mineração de dados pode ser usada junto com um Data Warehouse para ajudar com certos tipos de decisões. Para bancos de dados muito grandes, que rodam terabytes ou até petabytes de dados, o uso bem-sucedido das aplicações de mineração de dados dependerá, em geral, da construção de um Data Warehouse.

Gabarito: Errado

95.(CESPE / SERPRO – 2008)

A etapa de Mineração de Dados (DM – Data Mining) tem como objetivo buscar efetivamente o conhecimento no contexto da aplicação de KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Base de Dados). Alguns autores referem-se à Mineração de Dados e à Descoberta de Conhecimento em Base de Dados como sendo sinônimos. Na etapa de Mineração de Dados são definidos os algoritmos e/ou técnicas que serão utilizados para resolver o problema apresentado. Podem ser usados Redes Neurais, Algoritmo Genéticos, Modelos Estatísticos e Probabilísticos, entre outros, sendo que esta escolha irá depender do tipo de tarefa de KDD que será realizado. “Uma dessas tarefas compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores reais”. Trata-se de:

- a) Regressão.
- b) Sumarização.
- c) Agrupamento.
- d) Detecção de desvios.

Comentários:

Quem busca mapear registros de um banco de dados em um intervalo de valores reais é a Regressão.

Gabarito: Letra A

#vamosPraticar#

#VamosPraticar#

Questões Comentadas – Diversas Bancas

144.(ESAF / STN – 2018)

Uma técnica de classificação em Mineração de Dados é uma abordagem sistemática para:

- a) construção de controles de ordenação a partir de um conjunto de acessos.
- b) construção de modelos de classificação a partir de um conjunto de dados de entrada.
- c) construção de modelos de dados a partir de um conjunto de algoritmos.
- d) construção de controles de ordenação independentes dos dados de entrada.
- e) construção de modelos de sistemas de acesso a partir de um conjunto de algoritmos.

Comentários:

Uma técnica de classificação em Mineração de Dados é uma abordagem sistemática para a construção de modelos de classificação a partir de um conjunto de dados de entrada.

Gabarito: Letra B

150.(FUNDEP / IFN/MG – 2014)

Ao se utilizar a técnica de data mining (mineração de dados), como é conhecido o resultado dessa mineração, em que, por exemplo, se um cliente compra equipamento de vídeo, ele pode também comprar outros equipamentos eletrônicos?

- a) Regras de associação
- b) Padrões sequenciais
- c) Árvores de classificação
- d) Padrões de aquisição

Comentários:

O resultado de uma mineração de dados como o exemplo da questão é uma Regra de Associação.

Gabarito: Letra A

153.CCV-UFC / UFC / 2013)

Sobre Mineração de Dados, assinale a alternativa correta.

- a) É uma técnica de organização de grandes volumes de dados.
- b) É um conjunto de técnicas avançadas para busca de dados complexos.
- c) É o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida.
- d) É um processo automatizado para a recuperação de informações caracterizadas por registros com grande quantidade de atributos.
- e) É um processo de geração de conhecimento que acontece durante o projeto de banco de dados. Os requisitos dos usuários são analisados e minerados para gerar as abstrações que finalmente são representadas em um modelo de dados.

Comentários:

(a) Errado, não se trata de organização de grandes volumes de dados; (b) Errado, não se trata de busca de dados complexos; (c) Correto, pode ser definido como o processo de explorar grande quantidade de dados para extração não-trivial de informação implícita desconhecida – isto é, busca de insights em uma grande quantidade de dados; (d) Errado, não se trata de um processo automatizado, mas semi-automatizado – além disso, não se trata de um processo de recuperação de informações, mas de descobertas de informações; (e) Errado, na verdade ele faz parte de um processo de geração de conhecimento, sendo uma de suas fases.

Gabarito: Letra C

#VamosPraticar#

#VamosPraticar#

163.(FUMARC / PRODEMG – 2011)

Analise as afirmativas abaixo em relação às técnicas de mineração de dados.

- I. Regras de associação podem ser usadas, por exemplo, para determinar, quando um cliente compra um produto X, ele provavelmente também irá comprar um produto Y.
- II. Classificação é uma técnica de aprendizado supervisionado, no qual se usa um conjunto de dados de treinamento para aprender um modelo e classificar novos dados.
- III. Agrupamento é uma técnica de aprendizado supervisionado que particiona um conjunto de dados em grupos.

Assinale a alternativa VERDADEIRA:

- a) Apenas as afirmativas I e II estão corretas.
- b) Apenas as afirmativas I e III estão corretas.

- c) Apenas as afirmativas II e III estão corretas.
- d) Todas as afirmativas estão corretas.

Comentários:

(I) Correto. As regras de associação consistem em identificar fatos que possam ser direta ou indiretamente associados; (II) Correto. Trata-se de uma técnica de aprendizado supervisionado, isto é, as classes são pré-definidas antes da análise dos resultados. Além disso, essa técnica realmente usa um conjunto de dados de treinamento para aprender um modelo e classificar novos dados; (III) Errado. Agrupamento é uma técnica de aprendizado não-supervisionado.

Gabarito: Letra A

164. (FMP CONCURSOS / TCE/RS – 2011 – Letra B)

Mineração de Dados é parte de um processo maior de pesquisa chamado de Busca de Conhecimento em Banco de Dados (KDD).

Comentários:

Ele realmente é parte de um processo maior de pesquisa chamado de busca ou descoberta de conhecimento em bancos de dados.

Gabarito: Correto

166.(ESAF / MPOG – 2010)

Mineração de Dados:

- a) é uma forma de busca sequencial de dados em arquivos.

- b) é o processo de programação de todos os relacionamentos e algoritmos existentes nas bases de dados.
- c) por ser feita com métodos compiladores, método das redes neurais e método dos algoritmos gerativos.
- d) engloba as tarefas de mapeamento, inicialização e clusterização.
- e) engloba as tarefas de classificação, regressão e clusterização.

Comentários:

Nenhum dos itens faz qualquer sentido, exceto o último. A mineração de dados realmente engloba as tarefas de classificação, regressão e clusterização.

Gabarito: Letra E

#VamosPraticar#

Referências e links deste capítulo

1 <http://www.instagram.com/professordiegocarvalho>

2 <https://pt.wikipedia.org/wiki/Dados>

3

4 Data Warehouse (Armazém de Dados) é um repositório central de dados de várias fontes, projetado para relatórios e análises. Ele armazena dados atuais e históricos em um único local, criado pela integração de dados de várias fontes e processos e são projetados para dar suporte às decisões de negócios, permitindo que os dados sejam analisados de várias perspectivas. Eles também são usados para fornecer aos usuários uma visão abrangente dos dados, facilitando a análise de conjuntos complexos de informações.

5 No aprendizado de máquina, ruídos são flutuações aleatórias ou erros em conjuntos de dados que podem afetar a precisão de um modelo. O ruído pode ser causado por uma variedade de fatores, como erros de sensores, outliers, entrada incorreta de dados, etc.

6 <https://monkeylearn.com/word-cloud/>

7 Em tese, existe também o aprendizado semi-supervisionado em que parte dos dados são rotulados e parte não são rotulados (isso ajudar a reduzir custos do processo supervisionado).

- 8** Dica que eu encontrei recentemente: começou com A é uma técnica de aprendizado não supervisionado (
- 9** Também chamada de Detecção de Novidades, Detecção de Ruídos, Detecção de Desvios, Detecção de Falhas, Detecção de Exceções ou Mineração de Exceções.
- 10** Quem aí já ouviu falar no livro O Poder do Hábito?
- 11** Quando o k-NN é utilizado para classificação, contabiliza-se a classificação mais comum dos vizinhos mais próximos para estimar a classificação do novo dado; quando ele é utilizado para regressão, contabiliza-se a média dos valores dos vizinhos mais próximos para estimar a classificação.
- 12** Regras de Classificação são regras de fácil interpretação no estilo se (antecedente), então (consequente) em que o antecedente é um conjunto de testes e o consequente é a classe ou distribuição de probabilidades sobre as classes.
- 13** No exemplo, colocamos apenas algumas palavras no gráfico para facilitar a explicação didática, mas é óbvio que e-mails possuem outras palavras.
- 14** Quando não há ocorrências de uma palavra nos dados de treinamento (Ex: Relatório = 0), é necessário utilizar técnicas de suavização para evitar que o resultado total seja 0. Uma das técnicas mais utilizadas é a Estimção de Laplace, que soma uma unidade para todas as frequências.
- 15** Caso estivéssemos usando regressão, seria um Support Vector Regression (SVR).
- 16** Pode-se dizer também hiperplano unidimensional, hiperplano bidimensional e hiperplano tridimensional respectivamente.
- 17** Com uma exceção: quando estamos tratando de SVM Linear. Nesse caso, temos uma função de mapeamento linear.
- 18** <http://www.mercadolivre.com.br>
- 19** Apesar disso, não há garantia de que um ótimo global será atingido, ou seja, de que o melhor particionamento possível dos dados será encontrado.
- 20** No caso do k-médias, designa-se centroide do grupo; no caso do k-medoides, designa-se medoide do grupo.

21 Grafo é uma estrutura de dados que consiste em vértices ou nós conectados por arestas ou linhas, sendo utilizado para armazenar relações ou conexões entre objetos (Ex: redes sociais, redes de computadores, rotas de entrega, etc).

22 Essa é uma técnica de coocorrência conhecida como Análise de Cesta de Compra/Mercado, cujo objetivo é identificar combinações de itens que ocorrem com frequência significativa em bancos de dados e podem caracterizar, por exemplo, hábitos de consumo de clientes em um supermercado.

23 Também pode ser considerado um modelo de processos, um framework de processos, uma metodologia, entre outros.